nature

CO_2 STORAGE

Natural gas fields as a model for power-station emissions burial

Abstractions

LAST AUTHOR

Being able to read another person's mind is still science fiction. But Frank Tong, a cognitive neuroscientist at Vanderbilt University in Nashville, Tennessee, and his colleague Stephenie Harrison might have brought this fantasy a little closer to reality. Researchers thought that brain areas involved in the earliest stages of visual processing, including the primary visual cortex, could not retain the information they interpret from the signals received from the eye. Using functional magnetic resonance imaging (fMRI), Tong and Harrison have now shown that early visual areas do retain precise visual information about items that are no longer in the visual field — at least for a brief period (see page 632). Tong tells *Nature* more about the discovery.

What did you actually find?

We showed volunteers two striped patterns in different orientations and then asked them to remember one of the patterns for several seconds while we scanned their brains by fMRI — a technique that measures a signal produced by the increase in blood oxygenation that follows neural activity. By decoding the activity in the visual cortex, we could predict in more than 80% of the tests which of the two patterns a volunteer was remembering.

Were you surprised?

We thought we might find some evidence of visual memory in the visual cortex, but we were surprised to find it when brain activity was extremely low. It could be that when you're thinking about something, it is not at the same degree of vividness as when you are actually seeing it. Also, it could be that neurons in the visual cortex can transmit much information with little activity.

How were you able to interpret the signal?

Usually, fMRI signals are measured using 'voxels', a three-dimensional unit of measurement consisting of a few millimetres along each side. We used pattern analysis to pool the weak information contained in many individual voxels to obtain more robust information across the visual cortex. With this method, we can predict what people are seeing, paying attention to or actively remembering.

Will mind reading be possible some day?

We have a long way to go before these techniques could be applied to, say, a criminal investigation, but the possibility of reading out a person's thoughts does exist. But here we were reading out what our volunteers chose to remember, so people have some control over what thoughts can be read out. Right now, what we are doing is still fairly basic. ■

MAKING THE PAPER

Piergiorgio Picozza

An experiment to detect high-energy positrons pays off.

Seventy years ago, scientists first calculated that galaxies must contain additional, undetectable sources of mass — up to five times the mass of the detectable gas and stars. Piergiorgio Picozza, a physicist at the University of Rome Tor Vergata in Italy, has spent his career searching for this invisible 'dark matter', which is proposed as the source of the added mass, and he might now have found evidence for it.

Picozza has been investigating the formation of antimatter in space. Antimatter consists of particles that have the same mass as electrons and protons, but opposite properties such as charge. For example, the positively charged positron is the antimatter counterpart of the electron. Positrons can be produced by 'secondary processes', such as cosmic-ray nuclei smashing into interstellar dust, which occur at relatively low energies, but they might also arise directly from 'primary sources', such as dark-matter annihilations, that could generate positron–electron pairs at high energies. The latter process has not yet been confirmed. So a better understanding of positron formation could indicate the presence of dark matter. "A very important part of our job is to disentangle the sources of positrons," says Picozza.

To gather the necessary data, Picozza organized a collaboration of Russian, Italian, German and Swedish colleagues dubbed PAMELA — Payload for Antimatter–Matter Exploration and Light-nuclei Astrophysics. At first, PAMELA was difficult to get funded as a US-led collaboration had just begun similar work. But Picozza persevered and convinced European funders that two sets of data would be better than one. Specialized high-energy particle detectors to precisely measure the abundance of cosmic rays, electrons, positrons and other antimatter particles were sent into Earth orbit on board a satellite in 2006.

To identify possible primary source antimatter production, the team focused its analysis on the energy interval between 1.5 and 100 gigaelectron volts (GeV). If positrons are produced mainly from secondary sources, the ratio of positrons to electrons detected would be expected to decrease with increasing energy. But, surprisingly, the team found that this fraction increased significantly between 10 GeV and 100 GeV (page 607). The authors conclude that a primary source is needed to generate the high numbers detected at these higher energies.

Picozza is careful not to jump to the conclusion that their results prove that the primary source of antimatter is dark-matter annihilation. Pulsars, relics of massive stars that emit radiation, could also generate positrons. The ultimate confirmation that antimatter particles are produced from dark matter will come only if the Large Hadron Collider (LHC) at CERN near Geneva in Switzerland can experimentally produce 'dark matter particles'. "I remain open-minded about the possibilities, but if the LHC confirms our data, it would easily be the best result I — and more importantly, my young collaborators — will have achieved," says Picozza.

Until then, he hopes to take advantage of PAMELA's remaining time in space to follow antimatter production during a shift from low to high solar activity. The PAMELA data below 10 GeV were obtained in a period of low solar activity, and are remarkably different from previous data obtained during high activity. ■

FROM THE BLOGOSPHERE

Nature Chemistry (www.nature.com/nchem/index.html) has finally arrived! In a post in The Sceptical Chymist (http://tinyurl.com/c73cc8), associate editor Neil Withers announces the first issue, which is "freely available for everyone to read and (hopefully) enjoy".

Uppsala University postdoc and blogger Egon Willighagen has already taken a look (http://tinyurl.com/dfvgon). In his 19 March post, he happily notes that many of the papers have data-rich 'compound pages' in which readers can click on a compound number to view a full structure, with links to online databases.

In other papers, readers can click on the 'Show compounds' link that appears in the right-hand navigation panel and compound names in the text will be highlighted. Clicking these names reveals links to PubChem and ChemSpider.

Willighagen concludes that "*Nature Chemistry* really changes publishing of chemistry". In addition to the usual mix of research articles, reviews, News and Views and Research Highlights, the journal includes Blogroll, a quick overview of what has caught the editors' eyes in the blogosphere. ■

Visit Nautilus for regular news relevant to *Nature* authors ▸ http://blogs.nature.com/nautilus and see Peer-to-Peer for news for peer reviewers and about peer review ▸ http://blogs.nature.com/peer-to-peer.

nature

www.nature.com/nature

Vol 458 | Issue no. 7238 | 2 April 2009

Time for a concerted nuclear approach

Nuclear non-proliferation's moment has come. Scientists must help governments to seize a historic opportunity to avoid future apocalypses.

When leaders of the G20 nations gather in London this week, their attention will undoubtedly be focused on the current financial crisis. But it cannot be their exclusive focus: the crisis itself is a grim reminder that imminent global threats are best dealt with before the event, not after. And nothing poses a greater threat for creating further crises than nuclear weapons, either in existing stockpiles or through their acquisition by an increasing number of states — or by terrorists.

Fortunately, many of the G20 attendees seem to feel that urgency. Their host, UK prime minister Gordon Brown, has signalled that he is ready to put cuts to his country's arsenal on the table — although his government remains committed to a costly revamp of its deterrents, despite a lack of compelling justification. And US president Barack Obama and his Russian counterpart Dmitry Medvedev are expected to sign a pledge at the G20 meeting to reach an agreement by the end of the year to make substantial cuts to their nuclear arsenals.

This is excellent news, especially given how relations between the United States and Russia have soured over the past decade. The two countries first agreed to large reductions in their nuclear stockpiles under the Strategic Arms Reduction Treaty, which was formulated in 1982 and finally signed in 1991. But that treaty expires in December, and as yet no follow-up has been pursued. A new nuclear entente is sorely needed — not least to tackle the terrorist threat posed by the insecure stockpiles of weapons and fuel across the countries of the former Soviet Union.

But the world's leaders need to go much further. Over the past decade the whole fabric of the nuclear non-proliferation regime has begun to unravel — notably through the failure to implement ways to strengthen the Nuclear Non-Proliferation Treaty, such as through a Comprehensive Nuclear Test Ban Treaty. The situation is now dire. North Korea, which tested a nuclear device in 2006, seems set to test an intercontinental ballistic missile within days. Pakistan, which is estimated to have dozens of nuclear warheads, is politically unstable. And Iran, according to many scientists, now has enough fuel-grade low-enriched uranium to convert into a bomb's worth of highly enriched uranium, should it choose to do so.

These challenges will only grow more acute if, as expected, nuclear power is revived around the world as a way to mitigate climate change. A solution is urgently needed to ensure that the fuel intended for civilian nuclear reactors, as well as the huge amount of waste they produce, is not diverted to military ends. Some radical solutions are already under discussion, such as bringing all fuel-production facilities under multinational control.

Forging a consensus on these matters will not be easy. But scientists and engineers can play a crucial part by redoubling their efforts to create informal scientific and diplomatic backchannels. Particularly notable in that context is a conference taking place on 17–20 April in the Hague: the 58th annual meeting of the international Pugwash movement (see page 575). The movement's frequent convocations of influential scientists, politicians and other figures are credited with making key progress in arms control during the cold war. And although today's geopolitics are very different, the movement's efforts are as relevant as ever. Behind the scenes, for example, Pugwash is pursuing informal contacts with Iran to find ways out of that crisis. Scientists are also engaging in disarmament in newer organizations such as the non-profit US Nuclear Threat Initiative, which is working to reduce nuclear threats by championing a multilateral fuel bank, and a clean-up of stocks of highly enriched uranium.

Indeed, there is cause for optimism on the nuclear front. Obama's pledge to work towards a world free of nuclear weapons seems sincere, and is galvanizing support for new multilateral efforts in non-proliferation. With quick action, moreover, there is still time to build enough political momentum and preparation to make substantial progress at next year's crucial review conference of the Nuclear Non-Proliferation Treaty. The United States could send a strong signal here by sending the Comprehensive Test Ban Treaty to the Senate for ratification — as Obama has said he intends to do. As Brown said in a landmark speech on the topic on 17 March, it is time "to transform the discussion of nuclear disarmament from one of platitudes to one of hard commitments". ■

Clicking on a new chapter

The e-textbook is only one part of a bigger revolution in online learning.

For generations, students have flipped through their textbooks to amplify or clarify what they have heard in their lectures, to remind themselves how the various ideas relate one another, and — especially important in science courses — to find a good graphical depiction of the ideas they are struggling to understand. Once a student can picture in his or her mind the structure of DNA, say, or the mechanism of the greenhouse effect, much of the teacher's job is done.

Students will always need this kind of help; it is central to the learning process. But they might not be getting it from a printed textbook for much longer. The boundaries of the textbook have been stretching for some time now. Many already come with a CD attached, or include access to a website where updates and supplementary information can be found. Now those boundaries are threatening to burst entirely, as publishers experiment with making their textbooks available on personal computers, e-readers such as

the Amazon Kindle and handheld devices such as the iPhone (see page 568). The printed textbook will not vanish anytime soon — but a generation from now, it could be just a memory.

Yet at the same time, new technology is not limited to delivering the same type of content in new formats. E-textbooks are part of a much larger technological shift in the nature of teaching and learning. As is typical on the Internet, it is users who are driving some of the most popular innovations. Although the large publishing houses are understandably taking their time to consider how best to connect to new media, teachers and students, unconstrained by the need to protect jobs and revenues, are further ahead in experimenting with how to make the best use of virtual environments.

At the simplest level is the worldwide trend for both teachers and institutions to provide online access to course notes — often free of charge. Beyond that are collaborations between teachers to produce altogether new types of learning resource. At the University of Edinburgh, UK, for example, teachers have produced a set of free-to-download computer animations that illustrate concepts and phenomena in the physical sciences (see www.ph.ed.ac.uk/cgi-bin/interactive/applets).

And at a third level are virtual classrooms, in which teachers speak to global audiences through online classes and seminars, or via do-it-yourself online courses such as those offered by the US National Science Teachers Association in Arlington, Virginia. Indeed, more and more colleges and universities are taking courses almost completely online through 'virtual learning environments' such as the commercial Blackboard system, headquartered in Washington DC, or the open-source Dokeos platform from Europe. These environments not only allow students to access tests, homework, grades and lectures via the Internet, but they increasingly use wikis, blogs, messaging and even three-dimensional virtual environments such as Second Life to create online communities around each course. Such communities are particularly valuable for distance learning, to avoid students having to work in isolation.

"There is a ferment of creativity and innovation in education that deserves to be encouraged."

The result is a ferment of creativity and innovation in education that deserves to be encouraged. The funding agencies and private foundations are already doing so to some degree. The Edinburgh project, for example, was funded by Britain's Higher Education Academy, based in York. But they need to support such efforts more systematically — particularly by developing toolkits that make it easy for teachers to create instructional modules, and by encouraging the adoption of Sharable Content Object Reference Model and other such open standards for instructional software so that the modules can be used anywhere.

Textbook publishers would also do well to support such efforts, rather than ignoring or even resisting them, as the music industry tried to do with digital recordings. Textbooks were kings in a world where few other learning resources existed. University students, college libraries and school science departments had no option but to buy them. Now they have much more choice. ■

A bill against rights

Italy's Senate has approved a bill that ignores patients' wishes and the country's own constitution.

On 26 March, the Italian Senate approved a bill that would give physicians in the country the right to override the living wills of people who are in a persistent vegetative state, and to try to keep the patients alive through artificial nutrition.

The measure has caused intense controversy. Many countries have laws, or established codes of medical practice, that protect the expressed wishes of an individual to decline treatment if they become severely incapacitated and incapable of communicating. In most US states, for example, a doctor must negotiate with relatives via an ethics committee if he or she believes that a patient incapacitated in this way could benefit from additional treatment. The Italian bill, however, which is now being discussed in the lower house of parliament, the Chamber of Deputies, explicitly allows physicians to overrule such living wills. It also declares that artificial nutrition — which requires a feeding tube to be implanted into the stomach — is not a clinical intervention.

Curiously, the proposed law applies only to patients in the type of prolonged, deep coma known as a persistent vegetative state, and not to those with other, similarly incapacitating illnesses. This is because the bill has been prompted by the recent and much-publicized death of Eluana Englaro, who spent 17 years in a vegetative state after a car accident at the age of 21. Her father, arguing that his daughter had voiced a desire to be allowed to die if incapacitated, had pressed her reluctant doctors to cease artificial feeding. He eventually took legal action, winning in one court after the next in fighting off all the doctors' appeals. In February, he finally had her moved to a hospital that was prepared to remove the feeding tube. Prime Minister Silvio Berlusconi issued an emergency decree to block the process, but the Italian president refused to sign it. The constitutional crisis was averted when Englaro died on 9 February.

Surveys have indicated that a large majority of Italians do not support the idea that living wills could be ignored. But most relevant scientific societies have been quiet. The Federation of Italian Physicians published only a mild statement, after the Senate vote, suggesting that it should have been consulted.

As tragic as Englaro's situation was, media-fuelled emotion is not a good basis for lawmaking. The Italian constitution says that no one can be forced to undergo medical treatment without his or her approval. The Chamber of Deputies must now ensure that the bill is imbued with a suitable level of scientific and legal sophistication, and that it meets this constitutional provision. Discussion needs to embrace the requested wider consultation with the medical community and provisions should be made for care-givers' conscientious objection. But a physician whose conscience precludes his or her personally removing a feeding tube should not have the last say in the life or death of a patient whose wishes are clearly stated. ■

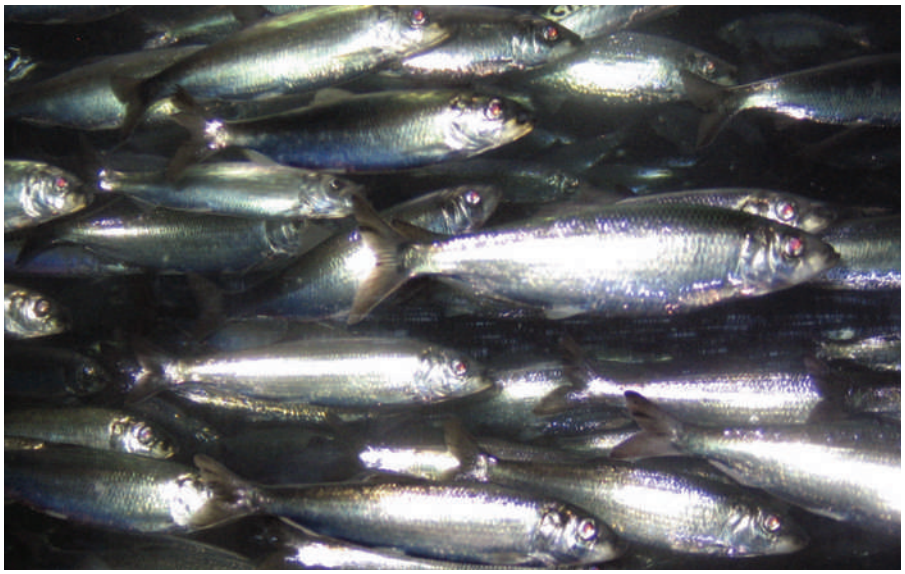
RESEARCH HIGHLIGHTS

School soundings

Science **323,** 1734–1737 (2009)

It is difficult to study what triggers shoaling in sea fish as the conglomerations can be tens of kilometres across and yet are still hard to find in the vast oceans. Nicholas Makris of the Massachusetts Institute of Technology and his colleagues have observed the genesis of an entire giant shoal for the first time, using a low-frequency acoustic technique that can take snapshots of areas up to 100 kilometres across every 75 seconds.

They found that spawning Atlantic herring (*Clupea harengus*) around the Georges Bank in the Gulf of Maine had to reach a critical density of 0.2 fish per square metre to trigger a rapid transition from anarchy to synchronization. After this transition the fish then proceed to migrate in their millions under the influence of a small number of leader fish.



H.BAESEMANN/DPA/CORBIS

MITOCHONDRIAL GENOMICS

Bloody anomaly

Genome Res. doi:10.1101/gr.083188.108 (2009)

Blood-sucking lice are common. Genetically, they are also unusual, say Renfu Shao at the University of Queensland, Australia, and his colleagues. Using information from the Human Body Louse Genome Project, the team found that the mitochondrial genome of the human body louse (*Pediculus humanus*) is splintered into 18 mini-chromosomes.

Chromosome fragmentation seems to have evolved along with blood sucking: the authors found it in human head and pubic lice, as well as in blood-sucking lice of other primates, but not in related lice that feed on other material. The chromosomal break-up may have been advantageous by increasing recombination between mini-chromosomes and introducing genetic variation that helped lice adapt to a bloody mammalian diet.

MECHANOCHEMISTRY

Tug of war

Nature Nanotechnol. doi:10.1038/nnano.2009.55 (2009)

Even the strongest molecular bonds break if yanked hard enough. But studying this effect requires a delicate tugging mechanism that can focus force controllably on individual bonds.

Roman Boulatov and his colleagues at the University of Illinois in Urbana-Champaign have found such a device: a rigid U-shaped molecule, stiff stilbene (pictured right), the ends of which are attached to the molecule under interrogation. Stilbene twists into a strained shape on exposure to light, pulling on its attached molecule. The force generated can be calculated from quantum mechanical principles, and altered incrementally depending on the length of an adjustable linker.

The researchers confirm a direct relationship between the force their probe exerts on a cyclobutene molecule and the rate at which a central bond falls apart.

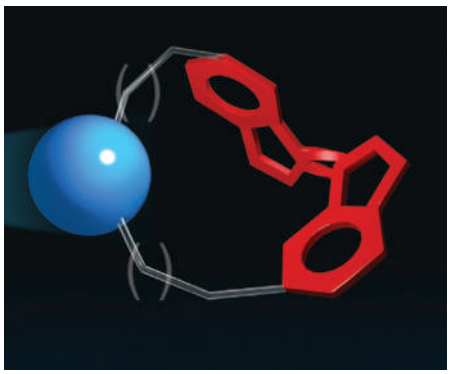
TRIBOLOGY

Brushing problems aside

Science **323,** 1698–1701 (2009)

The joints in human elbows, knees and the like exhibit very little friction even at moderately high pressure — man-made materials can offer nothing as good. Zwitterions might put that right.

Zwitterions are molecules with discrete positive and negative charges in different places. Jacob Klein of the University of Oxford, UK, and his colleagues have created polymer 'brushes' made of zwitterionic phosphorylcholine, in which the multiple positive and negative charges strongly attract water molecules, and attached them firmly to mica surfaces. The result is a system with very low friction when the surfaces move against each other, probably because the water molecules clinging to the phosphorylcholines prevent the brushes becoming entangled. The bound water can exchange freely with other water molecules, which also reduces friction.



This work might have application in biomedical devices where friction is often a problem.

ASTRONOMY

Slow revolution

Astrophys. J. **694,** 130–143 (2009)

Galactic archaeologists have identified a component of the Milky Way's halo that had been predicted but not seen before. The team, led by Heather Morrison at Case Western Reserve University in Cleveland, sifted through stellar velocity data from surveys going back to 1994, and found a group of stars marching to a different beat from the halo's original inhabitants. These stars were probably part of the outer halo and seem to have arrived at their positions more recently.

Some astronomers had theorized that the halo of stars centred on the Milky Way should contain two components. One, roughly spherical, would not rotate. The other, observed now for the first time, flattened into a thick, slowly rotating disk after the Galaxy's formation when stars from the outer halo drifted inwards.

This new component contains stars with eccentric orbits not found in the rapidly rotating main disk.

MARINE BIOLOGY

Deep-sea Methuselahs

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0810875106 (2009)

The longevity of deep-sea corals has been much debated: radiocarbon dating provides estimates of millennia, but counting growth rings gives ages of only a few hundred years. Brendan Roark at Texas A&M University in College Station, an advocate of the radiocarbon approach, now reports with his colleagues more evidence for extremely long-lived corals.

They show that, in some cases at least, the organic carbon that is acquired by the corals is 'fresh'. It is carbon rapidly transported from the surface ocean to the depths at which the corals live, rather than old sea-floor carbon in which the radioactive carbon-14 has already decayed.

The fresh diet means that the carbon-14 levels in the corals should accurately reflect their ages. On this basis the team estimates members of the black-coral genus *Leiopathes* to be 4,265 years old.

CHEMISTRY

Chemical scissors

Nature Chem. doi:10.1038/nchem.162 (2009)

A synthetic catalyst that mimics the chemical scissors at the heart of bacterial methane digestion can snap strong carbon–hydrogen bonds.

Previous attempts to copy the natural catalyst, which relies on a pair of iron atoms for its activity, produced catalysts that could only tackle relatively weak C–H bonds. The latest version, from Eckard Münck at Carnegie Mellon University in Pittsburgh and his colleagues, works thousands of times faster and breaks the toughest of C–H bonds, such as those in cyclohexane. It picks up electrons supplied by an electric current, and delivers them to the bond to prise the carbon and hydrogen atoms apart.

Although the synthetic di-iron catalyst does not match that of bacteria for speed, it goes one better by being able to break even stronger oxygen–hydrogen bonds.

CLIMATE CHANGE

Much travelled dust

Nature Geosci. doi:10.1038/ngeo474 (2009)

During the ice ages there was much more dust in the air over Antarctica than there is now, but its supply was sometimes rapidly curtailed.

David Sugden of the University of Edinburgh, UK, and his colleagues suggest that an 80,000-year record of the extent of the glaciers in Patagonia, the likely source of the dust, may explain the uneven pattern of dust deposition seen in Antarctic ice cores.

When the glaciers were extended, their sediment-rich discharge flowed out over extensive plains. Here, their dusty sediments would have been easily mobilized by the wind. When the glaciers retreated — as they did on occasion, even in an ice age — they discharged instead into lakes (pictured below), where the sediments simply accumulated. Glacier fluctuations correlate well with the Antarctic dust record.

B. HARRINGTON III/CORBIS

ECOLOGY

Saving songbirds

Ecol. Appl. **19,** 505–514 (2009)

The number of birds killed by crashing into communication towers could be reduced by about 50–70% by simply changing the towers' lighting systems, researchers say.

Millions of night-migrating songbirds collide with these towers each year. Joelle Gehring of Michigan State University in Lansing and her colleagues counted bird carcasses below 21 similar-sized towers in Michigan during two 20-day migration periods in 2005.

Towers with only flashing lights had a mean of 3.7 bird kills per season, whereas towers with both flashing and steadily burning lights had a mean of 13.

As the steady light may attract birds, the team suggests that tower operators turn off those lights or reprogram them to flash.

JOURNAL CLUB

Anthony J. Ryan
University of Sheffield, UK

A chemist welcomes an ingenious advance in plastics technology.

It's a rare joy to come across a communication that is truly concise, with a genuinely surprising but ultimately logical result, and compellingly modest conclusions that could materially benefit our society. Anne Hiltner at Case Western Reserve University in Cleveland, Ohio, and her colleagues take two well established facts — confined polymers form single crystals, and a blend of polymers, when stretched and folded by clever processing, makes very many thin layers — and use them to make something novel: a two-polymer blend with an oxygen permeability 100 times lower than either of its components (H. Wang *et al. Science* **323,** 757–760; 2009).

Plastics are often used in packaging as multilayer coatings. When each layer is thick, the barrier to oxygen is the sum of the properties of its components. The team found that as the layers were stretched, making them thinner, and folded back on themselves to make many layers, the plastic film became an even better oxygen barrier.

When a polymer crystallizes in a confined film it typically makes large pancake-like crystals around 10 nanometres thick and many micrometres across. Using simple mathematical models, the team showed that the improved barrier properties were due to the stretched and folded polymers forming alternating layers of such crystals. The core of each crystal is essentially impermeable to oxygen, which thus has to go across the pancake to find the edge — and at each alternate layer it faces another impermeable core: like a person having to go 1 kilometre sideways to go 1 metre forwards.

This astounding improvement is essentially free and could be incorporated into current packaging materials at little cost, reducing their environmental and energy impact. It makes a cold beer in a biodegradable plastic bottle a distinct possibility — and for me that would be a rare joy indeed!

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

Viral outbreak in China tests government efforts

Researchers call for greater focus on surveillance and genomics.

China has seen several outbreaks of hand, foot and mouth virus in recent years.

AP PHOTO

An outbreak of hand, foot and mouth disease in China, which since January has killed 19 children and made nearly 42,000 ill, has researchers calling for a better surveillance system to detect the disease and for action to speed up vaccine development.

"The situation of preventing and containing hand, foot and mouth disease is very serious at the moment," Deng Haihua, spokesman for China's health ministry, said last week. More cases are expected, as the disease normally peaks between May and July. In the absence of a drug treatment, the ministry is focusing on prevention and containment.

The outbreak is the latest in a series to have hit China in recent years, caused by a fast-spreading virus called enterovirus 71. "The persistence of enterovirus 71 outbreaks in China is a wake-up call," says Jane Cardosa, a virologist at the University Malaysia Sarawak in Kota Samarahan. In 1997, Sarawak saw the first outbreak of hand, foot and mouth disease in the Asia-Pacific region.

The disease causes flu-like symptoms, along with rashes on the hands and feet, and mouth ulcers. It can be caused by many types of human enterovirus belonging to the Picornaviridae family, which are mainly transmitted through faecal or oral routes. Although normally mild, the disease can be life-threatening: some viruses, particularly enterovirus 71, can cause inflammation of the brain stem, resulting in heart failure and fluid accumulation in the lungs.

In 1997 in Sarawak, more than 2,600 cases of the disease were reported and 29 people died. The next year in Taiwan, there were 129,000 reported cases and 78 deaths. In mainland China, the first reported case was in Shenzhen, Guangdong province, in 1999. At first, outbreaks were local and there were no reported fatalities (L. Li *et al. J. Clin. Microbiol.* **43,** 3835–3839; 2005). But since 2004, the outbreaks have become more severe and widespread, says Xu Wenbo, an infectious-disease expert at the Beijing-based China Center for Disease Control and Prevention.

Australian cap-and-trade plan comes under fire

The Australian government's proposed cap-and-trade scheme to regulate greenhouse gases, released in draft legislation last month, is facing mounting criticism from opposition politicians. Prime Minister Kevin Rudd, whose Labor party holds a slim majority in the House of Representatives and none in the Senate, is under pressure to alter the plan or risk defaulting on a promise to implement a system by 2010.

B. BAKKARA/AP

Australian climate-change minister Penny Wong.

Opposition leader Malcolm Turnbull of the Liberal party has called the scheme "irresponsible", and says it will cost jobs in a time of economic crisis. Meanwhile, the left-leaning Greens party argues that the emissions-reduction target, of 5–15% below 2000 levels by 2020, is "worse than useless".

Australia produces less than 2% of the world's greenhouse gases, but its per-capita emissions are among the highest in the world and rising (see chart). Decisive action from Australia could help build momentum for international climate-change negotiations in Copenhagen this December, says Senator Penny Wong, Australia's minister for climate change and water, who spoke on 30 March in Washington DC at a talk hosted by the Pew Center on Global Climate Change, based in Arlington, Virginia. "The best chance of an agreement at Copenhagen is for as many countries as possible to act," she says. "Australia is one of those."

"The best chance of an agreement at Copenhagen is for as many countries as possible to act."

In November 2007, a wave of public concern about climate in drought-ridden Australia helped Rudd win office over incumbent John Howard. On 10 March 2009, his government released draft legislation of an emissions-trading scheme that would begin on 1 July 2010.

Under the proposal, the roughly 1,000 Australian companies that emit 25,000 or more tonnes of carbon dioxide per year or the equivalent in other greenhouse gases would be required to obtain permits to emit, which could be bought at government auctions or traded. The country's total emissions would be controlled by a cap intended to achieve reductions by 2020 of at least 5% — up to 15% if other nations agree to similar targets — with a long-term goal of a 60% reduction below 2000 levels by 2050.

In May 2008, the country's health ministry added hand, foot and mouth to its category 'C' of notifiable diseases, meaning that all diagnosed cases must be reported through a national web-based system for disease surveillance, and took measures to streamline reporting requirements. The ministry also vowed to take a tough stance against cover-ups and last month sacked four health officials in Henan province for concealing the number of infections and deaths.

This year, enterovirus 71 has caused nearly all of the laboratory-confirmed cases in two hot-spots, the provinces of Henan and Shandong. Xu suspects that the disease's increasing virulence may be due to a genetic change in the circulating virus strain. Before 2004, the predominant strain was called C4b; since then, a different strain, C4a, has been most common (Y. Zhang *et al. J. Clin. Virol.* **44,** 262–267; 2009).

"The persistence of enterovirus 71 outbreaks in China is a wake-up call."

What caused this switch isn't clear, says Xu, as little is known about the genetics and transmission trends of the fast-mutating virus. Most studies have been clinical, aimed at, for example, identifying the strains behind a given outbreak and the disease's clinical features, especially when there are neurological complications. Many researchers say it is time to step up efforts to understand the basic biology of enterovirus 71 to speed vaccine development.

In a major push financed by the Chinese health ministry and the Center for Disease Control and Prevention, Xu and his colleagues measured the infection rate in adults and children during last year's outbreak and analysed stool samples and throat swabs taken from more than 18,000 patients. Preliminary results suggest that the infection rate is alarmingly high, meaning that there are large populations of virus carriers who do not show any symptoms of the disease.

Experts are divided as to how worried the world should be about the virus. Tom Solomon, a neurologist at the University of Liverpool, UK, argues that enterovirus 71 infection is underappreciated on a global scale and may pose a bigger risk to public health than is currently thought. But Hans Troedsson, the World Health Organization's representative in China, says "there is no cause for alarm". The public-health impact of hand, foot and mouth disease, including cases caused by enterovirus 71, is no more serious than other common childhood diseases, he says.

Troedsson thinks that the recent apparent increase in enterovirus 71 infection might be due to higher reporting rates rather than an increase in disease prevalence. "We will closely monitor the situation and decide polices accordingly," he says. ■

Jane Qiu

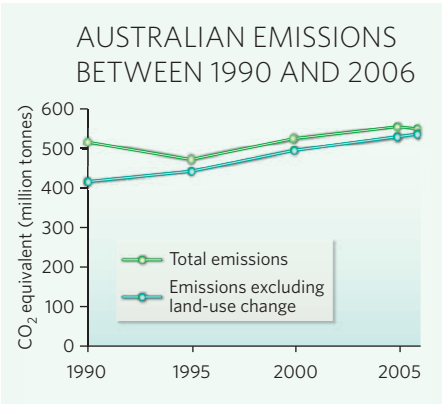
CORNELL UNIV.

Q&A: STEVE SQUYRES
Solar System decadal survey chair talks priorities.
www.nature.com/news

The plan also offers assistance to certain industries, which some opponents argue is too generous. As outlined in a government white paper released in December, emissions-intensive industries vulnerable to trade competition would get 60–90% free permits in the first year, and coal-fired power generators would receive an estimated Aus$3.9 billion (US$2.7 billion) in assistance over 5 years. Agriculture and deforestation, which account for about 27% of Australia's emissions, would not initially be included.

Two Senate committees are due to deliver reports reviewing the proposed scheme this month and next, and the government hopes to push the legislation through parliament by June. For the bill to pass, Rudd will need support from the Coalition, made up of the Liberal and National parties, or the Greens, says Andrew Macintosh, associate director of the Australian National University's Centre for Climate Law and Policy in Canberra. Although some industry representatives have opposed the bill, he says, others "recognize that this is still a very good deal" and could pressure the Liberals to accept it.

But the possibility of delays has raised concerns that companies will be rushed into auctions if the bill passes with a July 2010 timetable, says Brian Fisher, chief executive of consulting firm Concept Economics in Canberra, who worked on climate policy for the Howard administration. "Everybody's now panicking that they won't have time to see how this thing will work before they're forced to buy their first permits," he says. The government should set an initial low ceiling on permit prices to test the system and protect export industries, he says.



SOURCE: UNFCCC NATIONAL GREENHOUSE GAS INVENTORY DATA FOR THE PERIOD 1990–2006

Turnbull has argued that Australia should not finalize a scheme until after the negotiations at Copenhagen and after the United States reveals its plans. The latter came a step closer this week, as the US House of Representatives energy and commerce committee was set to release a draft cap-and-trade bill as *Nature* went to press. And on 20 March, the US Environmental Protection Agency submitted a proposed finding to the White House, widely thought to state that the greenhouse gases are pollutants endangering the public's health. Australia's experiences wrestling with cap-and-trade design issues could provide useful lessons for the United States as it formulates its own system, says Eileen Claussen, president of the Pew Center on Global Climate Change.

In Australia, the recent heat wave, wildfires and floods point to a need for urgent action, says Chris Cocklin, a sustainability policy expert at James Cook University in Townsville. "Every year we wait," he says, "it's just too damn long." ■

Roberta Kwok

Congress probes NIH stimulus funds

Scrutiny also aimed at National Children's Study.

Hard on the heels of their $10.4-billion gift to the US National Institutes of Health (NIH), members of Congress have made it clear that they will keep close tabs on how the biomedical agency spends the money.

At a 26 March hearing of the House subcommittee that funds the $30.3-billion agency, Democrats and Republicans grilled top NIH officials on how — and by how much — it will boost the economy with the $10.4 billion supplied in February's economic stimulus package (see *Nature* **457**, 942–945; 2009). "With that kind of increase, the committee will be watching carefully to be sure that the NIH spends it in a way that both stimulates the science [and creates] high paying jobs across the country," said Jesse Jackson Jr (Democrat, Illinois), who chaired the hearing.

Jackson said he was concerned that the abrupt curtailment of the bolus of funds, on 30 September 2010, could leave many scientists stranded in a much more difficult funding environment. "The Recovery Act funding is a double-edged sword," he said. "The prosperity is short-lived."

Raynard Kington, acting NIH director, told the subcommittee that the agency aims to avoid repeating the "not so soft" landing that occurred after its budget doubled between 1998 and 2003, then plateaued. The short-term projects slated for the NIH's stimulus money (see 'Spending power') include a new category announced last week: 'grand opportunities' grants, aimed at large-scale projects costing more than $500,000 per year. The NIH intends to fund these at about $200 million in total.

SPENDING POWER

The NIH received $10.4 billion in the US economic stimulus package. $7.4 billion of that will be used by individual institutes, much of it on grants already in the applications pipeline. The new pot of money many scientists are scrambling for is the $800 million granted to the director's office. Of this, roughly $291 million has been committed:

- At least $200 million to Challenge Grants
- Roughly $91 million to programmes including Signature Initiatives, Core Centers for Enhancing Research Capacity in US Academic Institutions, and summer training programmes

The remaining roughly $509 million will be committed later by the director, partly in Grand Opportunity Program grants.

However, Kington concedes, the agency may see a rise in grant applications beginning in 2011 if the stimulus money works as hoped to foster new discoveries and accelerate research. If that happens, he says, "we believe the success rate may drop at least several points from what it has been if we don't have a substantial increase in our budget".

Republicans on the subcommittee seemed sceptical that the stimulus funding would be spent on the best science. "Give us some confidence that, one, this will stimulate the economy as intended and, two, that you are not just going to be throwing money at new projects that hadn't made the [fundable] list before," said Dennis Rehberg (Republican, Montana). The agency is revisiting 14,000 grant applications already in the NIH's pipeline that were judged to be scientifically meritorious but that were not funded in its last round of reviews. These may now receive funding if they can show potential to make progress within two years.

Kington defended those applications, calling the projects at "the top, right below our funding level". As for job creation, he says, the biomedical agency has estimated that each of its grants supports on average "six or seven jobs in part or in full". Pressed further by Rehberg, Kington said he would get back to him with the exact number of jobs to be created with the NIH's $10.4 billion. Few expect this to be an easy number to find (see page 563).

Also under fire at the same hearing was the National Children's Study, a project first authorized by Congress in 2000 that aims to follow environmental and genetic influences on the health of 105,000 children from the womb to the age of 21. During its planning phase, the Bush administration repeatedly tried to kill its funding, and Congress repeatedly restored it. This year, it is receiving $192 million — up from $111 million in 2008. Its seven 'vanguard' centres, conducting its pilot phase, will all be open and beginning to enrol patients this month.

But shifting numbers on its estimated cost have become a political pitfall. Several months ago, NIH officials called Todd Tiahrt, the subcommittee's senior Republican, to explain that the initial $3-billion-plus price tag on the study could end up being double that amount. Last week, Kington told Tiahrt that the agency had been estimating "a moving target" while the study was in the planning stages, and that when it became apparent that the study would cost more than originally expected, the NIH decided not to adjust the estimate upwards until the results of the pilot study were in. "That was an error in judgement," Kington says. "We have every plan to bring the cost down."

After the hearing, Duane Alexander, the director of the National Institute of Child Health and Human Development in Bethesda, Maryland, called the contention that the project's costs had doubled "a myth". During the pilot phase of the study, he says, "we have an attractive tree to hang ornaments on", referring to the lengthy list of subprojects encompassed by the pilot. "There was never any expectation or intent that we would be able to fund all that." ■

Meredith Wadman

See also Party of One, page 563.

NIH

The NIH must detail the economic benefits and number of new jobs it will create with stimulus monies.

HAVE YOUR SAY
Comment on any of our news stories, online.
www.nature.com/news

Sonar mapping ventures into uncharted waters

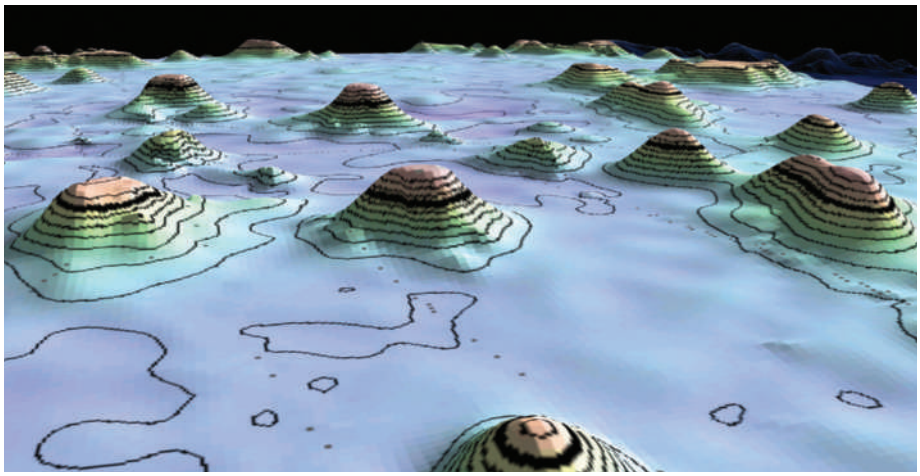
Ships cruising the globe may soon be able to help scientists to chart seamounts rising from the ocean floor.

Less than 1% of the 47,000 known seamounts standing taller than 500 metres have been mapped in detail. In 2005, the dangers this poses became clear when the nuclear submarine *USS San Francisco*, travelling submerged about 600 kilometres south of Guam, struck an uncharted seamount, damaging the vessel and killing a sailor.

A new system using a basic GPS device coupled to a computer would allow anything from freight ships to pleasure yachts carrying sonar to help chart seamounts, which could number as high as 200,000, oceanographers say.

The initiative is an outgrowth of the Seamounts '09 Workshop, held 19–21 March at the Scripps Institution of Oceanography in La Jolla, California. The idea is to take advantage of single-beam and multi-beam sonar now aboard many vessels. "This is a really cool opportunity to take the baby step to image these features," says meeting chairman Hubert Staudigel of Scripps.

Theoretically, any vessel could gather data from regions of interest, but the quality of the imaging depends on how deep the ship's echo-sounder can probe. The ocean has an average depth worldwide of about 4,000 metres; a typical navigation sonar reads only to 1,000 metres, but that means it still could pick up some tall seamounts.



P. WESSEL/D. SANDWELL

Mountains, mountains everywhere: seamounts are less well mapped than the volcanoes of Mars.

Government sonar data are typically hoarded for many years. The US Navy, for instance, is soon expected to release a massive cache of sonar survey data that it has gathered over the past few decades, says Christopher Fox, director of the National Geophysical Data Center in Boulder, Colorado; oceanographers hope that it will contain information on many unknown seamounts. Google has also been pushing for such data to be released to incorporate into its Google Ocean feature (see *Nature* **457,** 1065; 2009).

In the meantime, oceanographer David Sandwell of Scripps and his colleagues have created a program to allow anyone to engage in seamount mapping. Soon to be made available online (http://topex.ucsd.edu/marine_topo), the program allows people to superimpose the routes of research ships over ocean bathymetry data that indicate where seamounts may exist. Ships steaming near these huge unprobed regions could then send in their data for analysis.

The trick now is to create an easy way to access and store the data centrally. ■

Rex Dalton

Research review boards dogged by criticism

An undercover investigation into the system that regulates human experimentation in the United States has revealed flaws that expose it to 'unethical manipulation', the Government Accountability Office reported last week.

The federal inquiry was launched in January 2008 to probe the network of institutional review boards (IRBs) that oversee research using human patients, such as clinical trials. The boards are often run by universities and hospitals, but, with researchers clamouring for their proposals to be reviewed more quickly, a burgeoning industry of independent, for-profit IRBs has recently emerged.

In a hearing before the House Committee on Energy and Commerce on 26 March, government investigators reported that they had registered fictitious IRBs with the Office for Human Research Protections — including one called Phaké Medical Devices, supposedly based in 'Paynesville, South Carolina'.

At the hearing, protections office director Jerry Menikoff noted that the registration of IRBs is a simple listing process that does not involve background checks. That system was recommended following a previous inspection of the programme, he said.

Investigators also advertised a bogus IRB pledging "fast approvals guaranteed", and naming the fictitious board's president after a three-legged dog called Trooper. Six companies responded to the advert, and one attempted to hire the IRB.

In a separate arm of the inquiry, an intentionally vague research protocol was concocted and submitted to three real IRBs. The proposal called for a litre of a fake gel to be poured into the abdominal cavity of women to ease recovery after surgery. Two of the IRBs rejected the proposal outright, with one board member calling it "the riskiest thing I've ever seen", investigators reported. But one IRB approved the protocol unanimously.

"Our investigation showed the current system is highly vulnerable to unethical or incompetent actors," says Gregory Kutz, managing director of forensic audits and special investigations at the Government Accountability Office.

The report falls short of a comprehensive review of independent IRBs but is still valuable, says Trudo Lemmens, a bioethicist at the University of Toronto in Canada. "It's more or less anecdotal," he says, "but it confirms that there is a problem in how the system is constructed." ■

Heidi Ledford

Fungus farmers show way to new drugs

M. MOFFETT/FLPA

In a mutually beneficial symbiosis, leaf-cutting ants cultivate fungus gardens, providing both a safe home for the fungi and a food source for the ants. But this 50-million-year-old relationship also includes microbes that new research shows could help speed the quest to develop better antibiotics and biofuels.

Ten years ago, Cameron Currie, a microbial ecologist then at the University of Toronto in Ontario, Canada, discovered that leaf-cutting ants carry colonies of actinomycete bacteria on their bodies (C. R. Currie *et al. Nature* **398,** 701–704; 1999). The bacteria churn out an antibiotic that protects the ants' fungal crops from associated parasitic fungi (such as *Escovopsis*). On 29 March, Currie, Jon Clardy at the Harvard Medical School in Boston and their colleagues reported that they had isolated and purified one of these antifungals, which they named dentigerumycin, and that it is a chemical that has never been previously reported (D.-C. Oh *et al. Nature Chem. Bio.* doi: 10.1038/nchembio.159; 2009). The antifungal slowed the growth of a drug-resistant strain of the fungus *Candida albicans*, which causes yeast infections in people.

"These ants are walking pharmaceutical factories."

Because distinct ant species cultivate different fungal crops, which in turn fall prey to specialized parasites, researchers hope that they will learn how to make better antibiotics by studying how the bacteria have adapted to fight the parasite in an ancient evolutionary arms race. "These ants are walking pharmaceutical factories," says Currie, now at the University of Wisconsin, Madison.

That's not the end to the possible applications. The ant colonies are also miniature biofuel reactors, Currie reported on 25 March at the Genomics of Energy & Environment meeting at the Joint Genome Institute in Walnut Creek, California. Each year, ants from a single colony harvest up to 400 kilograms of leaves to feed their fungal partners. But no one has worked out how the fungi digest the leaves, because samples of fungus grown in petri dishes can't break down cellulose, a tough molecule found in plant cells. Researchers are keenly interested in better ways to break down cellulose, because it might allow them to make more efficient biofuels than those made from sugary foods, such as maize (corn).

So Currie and his colleagues sequenced small segments of DNA from bacteria and other organisms living in fungus gardens in three Panamanian leaf-cutting ant colonies. They then compared the DNA against databases to identify what species were living in the fungus gardens, and what genes they contained.

This 'metagenomics' approach found that there are many species of bacteria in the fungus gardens that are capable of breaking down cellulose. The team also detected the genetic signatures of fungal enzymes that can break down cellulose, which raises the question of why the fungi can't break down cellulose in the laboratory.

Currie suggests that the newfound bacterial and fungal enzymes might be efficient at digesting cellulose because they have evolved for centuries along with the ant-fungal symbiosis. This could mean that the fungus can only break down cellulose in its natural context, or that the enzymes Currie detected are brought into the colony from outside. "The idea is that the ants' long evolutionary history may help us in our own attempts to break down plant biomass," he says.

Dismissed researcher wins court battle

One of Germany's largest research centres was wrong to dismiss without notice one of its institute directors, a court in Munich has ruled.

The German Research Centre for Environmental Health in Munich had claimed that the dismissed scientist, immunologist Jean-Marie Buerstedde, was aggressive with colleagues and failed to nurture "relationships based on trust and respect" with students in his charge.

The centre provided the court with a long list of incidents involving Buerstedde, which describe him shouting insults, displaying insensitivity to students' personal difficulties, and forcing colleagues to work long hours and weekends. Doctoral students who complained about Buerstedde say that they were sometimes reduced to tears.

Speaking before the court ruling, Norbert Blum, the centre's chief financial officer, said: "We have special rules to protect young scientists, and Buerstedde broke them seriously."

Buerstedde says that his style of working was needed to remain at the forefront of the competitive field of antibody hypermutation. He brought a case of unfair dismissal against the centre after he was sacked without warning on 4 June 2008. He says that he was told to clear his desk and forbidden to enter the centre's grounds. His access to professional e-mail was cut off, he claims, making it difficult for him to complete research projects.

Blum insists that the centre had no intention of preventing Buerstedde finishing projects, adding that it had been "necessary to remove him from the scene so there would be no confrontation". He said that the centre's board had been "shocked by the extent of the problems caused by [Buerstedde]".

But the court ruled that the charges were not sufficiently well documented to assess the damage done by Buerstedde's alleged behaviour. It added that most of the charges did not justify dismissal without the normal warnings and meetings, and that the number of complaints alone was insufficient to dismiss Buerstedde without notice.

The centre said that owing to staff absences, it could not comment on the ruling before *Nature* went to press. Unless the centre appeals the decision before the end of

Studies of bacteria on leaf-cutting ants could yield new antibiotics.

Other researchers call Currie's findings interesting, but say they wanted to see a more thorough analysis of the data. "It's interesting that he found these fungal enzymes in the gardens that he didn't expect [based on] what the fungus was capable of doing by itself," says John Taylor, a mycologist at the University of California, Berkeley.

Taylor says that Currie's continued scrutiny of the lives of ants provides insights into the web of interactions necessary for the survival of any single species. "I think the coolest thing about this is that you start with one organism, and then you find more and more organisms involved in the relationship," he says. It may take a village to raise a child; it seems it also takes a village to break down cellulose. ■

Erika Check Hayden

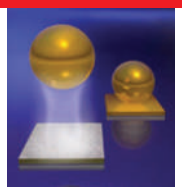
Visit http://tinyurl.com/ddh8o3 to see Cameron Currie discuss his research.

April, Buerstedde expects to return to work.

Buerstedde admits that he can be impatient, but says that many enjoy working with him. One colleague told *Nature* that Buerstedde needed to learn to hold his tongue, but that he was appalled at the peremptory dismissal. "I didn't think it was possible in a country like Germany that someone could be dismissed without being given a chance to hear the charges or defend himself against them," said the colleague, who did not wish to be identified.

Some of Buerstedde's external collaborators have also expressed dismay in open letters. "I was truly shocked that you … have been fired so abruptly," wrote David Schatz, an immunologist at Yale University in New Haven, Connecticut. "It is a blow to the integrity of the research process and to academic freedom." ■

Alison Abbott



J. PENNI & F. CAPASSO

SLIPPERY NANOMACHINES
Quantum stickiness need not slow tiny devices.
www.nature.com/news

Quark statistics shed light on Universe's symmetry

The fundamental asymmetry in the laws of physics called charge–parity (CP) violation is tiny, yet it looms large enough in physics to have led to Nobel prizes on three occasions. A persistent puzzle is why the asymmetry is so small — some theories imply that it could, and perhaps should, be much bigger. Now, research[1,2] is bolstering a previous suggestion that the smallness is not a mystery, but rather an inevitable consequence of another basic fact in physics: that the three known families of quarks have the masses that they do.

The findings, by Gary Gibbons and his colleagues at the University of Cambridge, UK, are spurring discussions about whether the laws of physics are 'fine-tuned' — that is, whether the magnitudes of various physical constants should be considered peculiarly unlikely. And they hint at the possibility of probing physics beyond the standard model, which describes all the known particles and forces at the subatomic scale.

CERN

Kaon decay: keeping track of the Universe.

CP violation means that some physical laws are altered if a subatomic particle is exchanged for its antiparticle, and at the same time left and right are swapped as in a mirror. The very subtle effect was first observed in 1964 in the way that exotic particles called neutral kaons decay. But some theories suggest that the asymmetry should be about a thousand times bigger than it is, leading scientists to wonder whether some unknown physical principle keeps the effect so small. Gibbons and his colleagues now suggest that the magnitude of the CP violation is just what should be expected given the observed masses of quarks (which make up protons, neutrons and other weighty particles).

The link between CP violation and quark mass is well known. In work that won them last year's physics Nobel, Japanese physicists Makoto Kobayashi and Toshihide Maskawa showed in 1973 that CP violations are inevitable if there are more than two types of quark (and corresponding antiquarks) that have different masses. Their finding connected CP violation to the hierarchy of quark masses, and suggested a way to work out its magnitude.

That shifted the question to why quark families have the masses they do — something not explained by the standard model, but that could fall out of a deeper, as yet unknown theory. In 1998, John Donoghue of the University of Massachusetts in Amherst suggested[3] that rather than predicting exact masses for specific quarks, such a theory might predict a 'landscape' of allowed masses, of which those observed in this Universe are typical examples.

In a 2006 paper[4], he and his co-workers showed that, by assuming a simple distribution of possible values for the quark masses, they could calculate the range of the most likely values of a parameter called *J* that quantifies CP violation. And they found the observed value fell squarely within that range. Gibbons and his colleagues now find much the same result when they assume a different statistical distribution of quark masses in the hypothetical landscape. "I can only speculate why the earlier work went largely unseen," says Max Tegmark of the Massachusetts Institute of Technology in Cambridge.

Tegmark says the work is "very interesting", but others disagree about its significance. Physicist Alexei Grinbaum of CEA-Saclay in Gif-sur-Yvette, France, says that the results do not dismiss the fine-tuning issue, but just shift the responsibility for fine-tuning to the hierarchy of quark masses. Particle physicist Graham Ross of the University of Oxford, UK, agrees with that, but also feels that the mass hierarchy can itself by explained by symmetry arguments — so there is no great mystery in the first place. ■

Philip Ball

1. Gibbons, G. W., Gielen, S., Pope, C. N. & Turok, N. *Phys. Rev. Lett.* **102,** 121802 (2009).
2. Gibbons, G. W., Gielen, S., Pope, C. N. & Turok, N. *Phys. Rev. D* **79,** 013009 (2009).
3. Donoghue, J. F. *Phys. Rev. D* **57,** 5499–5508 (1998).
4. Donoghue, J. F., Dutta, K. & Ross, A. *Phys. Rev. D* **73,** 113002 (2006).

K. CAMPBELL/GETTY

GOT A NEWS TIP?
Send any article ideas for *Nature*'s News section to **newstips@nature.com**

Retracted paper rattles Korean science

Authors disagree over work aimed at gene therapy for diabetes.

Nature this week is retracting a 2000 paper that promised an advance in diabetes treatment using gene therapy. Confusion surrounding the paper, including allegations about fraudulent data, continues to afflict the South Korean science community.

The paper's authors, led by Hyun Chul Lee of Yonsei University in Seoul, claimed to have created a treatment for type 1 diabetes, a condition in which the immune system destroys the insulin-producing cells needed to regulate glucose levels. Lee's team used a recombinant virus to introduce a gene for an insulin analogue into diabetic rats and mice, which was expressed in response to blood glucose levels and alleviated symptoms. The team suggested the treatment could be adapted for humans (H. C. Lee *et al. Nature* **408,** 483–488; 2000).

Now, having yet to repeat the experiment, Lee has asked *Nature* to retract the paper (see page 660). "I don't know the reason why the experiments are not reproducible," says Lee. He suggests that the original gene construct, pLPK-SIA — a combination of the virus vector, the insulin analogue and a promoter that regulates the expression of the analogue in response to glucose levels — might have mutated after the original experiment.

The background to the retraction is contentious. A researcher who joined the laboratory in 2001 tried and failed to initiate preclinical trials in bigger animals such as dogs and monkeys. But the researcher, who does not want to be identified for fear that acting as a whistleblower could harm his career, says he didn't find any pLPK-SIA in the laboratory, so with another researcher in the lab he tried to remake it according to the methods section from the original paper. Lacking essential ingredients, they eventually gave up.

The anonymous researcher says one of the paper's authors, Su-Jin Kim, who created the gene construct before moving to the University of Calgary in Canada, refused to send him samples. Kim says she deferred on this matter to her new boss, Ji-Won Yoon. The researcher, however, says that in e-mail exchanges, Yoon told him to ask Kim for samples. Yoon, also a co-author on the *Nature* paper, died in 2006.

Lee fired the anonymous researcher in August 2005, citing unhappiness with his work. Lee says that in 2008 the researcher threatened to disclose faults in the paper unless given money, grants and a new job. The researcher admits that he asked for a new position as compensation for losing what he calls four-and-a-half years trying to reproduce the results. He alleges that he was fired after advising Lee to retract the paper, which Lee denies.

M. DONNE/SPL

A retracted paper suggested that gene therapy could be used to treat type 1 diabetes.

In April 2008, Yonsei University started an investigation, chaired by chemist Won-Yong Lee. On 30 December the committee recommended a retraction based on multiple points, including the apparent duplication of figures and the fact that it could not confirm the key construct existed when the experiment was carried out. Won-Yong Lee says that the committee members examined Kim's lab notes and thesis, and alleges that "the duplication was more than a simple mistake", including the reuse of data as well as cutting, pasting and otherwise adjusting figures. In addition, "the pLPK-SIA found in the laboratory and deposited at a cell-line bank had mutations that would make the plasmid non-functional", Won-Yong Lee wrote in an e-mail to *Nature*'s news team.

The committee says that Kim and Yoon tried to reproduce the experiments; Kim, who is now at the University of British Columbia in Vancouver, says she did not, and didn't know there was a problem until last year. She says she has some of the pLPK-SIA and that the problems with figures were probably a mistake made when forwarding to colleagues, or in labelling. She faults the committee for choosing "to rely on the memory of witnesses who were testifying about experiments that took place 8–10 years ago". Kim refused to sign the retraction letter, calling the original experiment a success, based on lab notes. She also filed an injunction, currently under consideration in the Seoul District Court, to prevent the university releasing its full report.

Nature's policy is that it will permit retraction of a paper without the sign-on of all authors, while making clear which authors disagree with the retraction. A *Nature* spokesperson notes that underlying problems with a paper, if they exist, can be difficult to detect through standard peer review.

Won-Yong Lee says that the university ethics committee will decide whether any of the researchers involved will be censured after the court reaches a decision, expected within a few months, regarding the injunction.

The anonymous researcher faults the committee for, in his view, refusing to investigate several other alleged problems with the paper. Two months ago, he sent a letter of complaint to the Korea Research Foundation, which funded the research, but has yet to hear back. "Yonsei University investigated into the case in a way that generated minimal damage against Yonsei University," he comments. Won-Yong Lee disagrees strongly, saying that the committee had members from other institutions that had no vested interest in protecting Yonsei University or Hyun Chul Lee.

The researcher and Kim agree that reproducing the experiment would resolve the situation. Kim says she will ask her current boss to share pLPK-SIA samples with other researchers to do just that. ■

David Cyranoski

Climate experts urge G20 to make stimulus green

Climate-change analysts have urged leaders of the world's largest economies to invest more of their stimulus packages in reducing greenhouse-gas emissions.

Ottmar Edenhofer, co-chair of the Intergovernmental Panel on Climate Change, and Nicholas Stern, chair of the Grantham Research Institute on Climate Change and the Environment at the London School of Economics and Political Science, are aiming their report, *Towards a Global Green Recovery*, at politicians attending the G20 summit in London on 2 April.

The report estimates that almost $400 billion of the total $2,610 billion in economic-stimulus packages unveiled so far by the G20 nations has been earmarked for green measures such as renewable-energy projects (see chart). China says it will devote almost 35% of its stimulus spending (about $200 billion) on green projects in 2009 and 2010, and South Korea plans to devote more than 80% of its $38-billion stimulus on green measures in the next four years.

For more G20 coverage see www.nature.com/news.

SOURCE: TOWARDS A GLOBAL GREEN RECOVERY

Geometric work secures top maths prize

Mikhail Gromov won the 6-million-Norwegian-kroner (US$900,000) Abel Prize last week for his work on advanced forms of geometry. The Russian expatriate holds appointments at the Institute of Advanced Scientific Studies outside Paris and the Courant Institute of Mathematical Sciences at New York University. The Abel committee cited Gromov for his contributions to the study of Riemannian geometry, symplectic geometry and group theory.

Gromov is "renowned among mathematicians for his original approach", says Ian Stewart, a mathematician at the University of Warwick, UK, and his work has guided many other mathematicians and physicists. The Abel Prize was founded in 2003 by the Norwegian Academy of Science and Letters to complement the Nobel prizes, which do not reward work in pure mathematics.

For a longer version of this story, see http://tinyurl.com/abelprize.

Grazing limits effects of ocean fertilization

Preliminary results from a controversial Indo–German ocean fertilization experiment (LOHAFEX) have cast doubt on whether stimulating algal growth can help the sea sequester substantial amounts of carbon dioxide.

Earlier this year, researchers aboard the German research vessel *Polarstern* (pictured) poured 20 tonnes of iron sulphate over a 300-square-kilometre area of the Southern Ocean around the Antarctic (see *Nature* **457,** 243; 2009).

ALFRED-WEGENER-INST.

However, grazing by small crustaceans prevented blooms from growing as much as some had hoped, according to Germany's Alfred Wegener Institute for Polar and Marine Research in Bremerhaven, one of the experiment's backers. Furthermore, a lack of silicic acid in the water restricted the growth of diatom plankton, which are more resistant to predation than the algae. The fertilization therefore removed only a "modest amount" of carbon from the environment.

Drug patent pools start to take shape

GlaxoSmithKline, the world's second-largest pharmaceutical company in terms of sales, has fleshed out proposals outlined last month to create a pool for companies to share patents to boost research into neglected diseases (see *Nature* 457, 1064–1065; 2009).

The company says that it will put some 500 patents and 300 pending applications into the pool, and has confirmed that on 1 April it will cut the price of its drugs in the world's 50 poorest countries to no more than 25% of prices in the developed world.

On 24 March, Ivan Lewis, the UK minister for international development, called for other pharmaceutical companies to contribute to both GlaxoSmithKline's patent pool and another pool for AIDS drugs that is being established by UNITAID, an international organization that negotiates lower drug prices.

Gates supports Chinese tuberculosis drive

China this week announced new measures to tackle its growing problem with tuberculosis (TB). On 1 April, health minister Chen Zhu and Bill Gates announced a partnership, supported by a 5-year US$33-million grant from the Bill & Melinda Gates Foundation, to pilot new diagnostic tests, monitoring strategies and treatments for the disease. The Chinese government will scale up the most effective of these trials.

A day earlier, the Chinese Academy of Sciences and the Gates-supported Global Alliance for TB Drug Development signed a partnership to search for anti-TB drugs among Chinese herbal medicines.

The announcements came at the start of a three-day meeting in Beijing, organized by the World Health Organization, where health officials from 27 countries are discussing how to control multidrug-resistant TB.

Fossils protected in US land legislation

After nearly 20 years, US scientists have won approval for a law that seeks to protect vertebrate fossils found on federal lands.

The US Vertebrate Paleontological Resources Preservation Act was included in omnibus land-management legislation signed into law on 30 March by President Barack Obama.

The bill means a permit is needed to collect any scientifically significant vertebrate fossil, officials say. But it would allow 'casual collecting' of common fossils. Details of how the law will be applied are yet to be finalized.

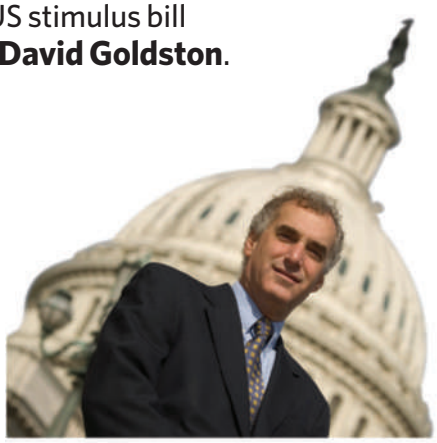
Officials at the Society of Vertebrate Paleontology have pushed for the legislation because of the widespread practice of commercial collecting, where important specimens may be sold and not recorded in the scientific literature.

Correction

The article 'Supplanting the old media?' (*Nature* **458,** 274–277; 2009) incorrectly stated the web traffic received by Derek Lowe's blog, In the Pipeline. The blog receives around 200,000 page views each month, not each week.

Mean what you say

Promises about job creation in the US stimulus bill may be coming home to roost, says **David Goldston**.



PARTY OF ONE

Scientists may be about to learn an important, and perhaps surprising, lesson about Washington: words matter. Rhetorical strategies crafted to push a particular bill affect expectations about the impact of that measure and can take on a life of their own. The stimulus package that became law in February, and will provide more than US$21 billion for research and development, is a case in point.

The Obama administration, Congress and advocacy groups sold the stimulus package to each other and to the public primarily as a way to create and retain jobs in the near future. The research funding in the bill was no exception, even though it was understood it could also promote longer-term economic growth. Congressman Rush Holt (Democrat, New Jersey), a physicist and a strong advocate for the research spending in the package, made the political linkage between research and jobs clear when he spoke at a conference on R&D priorities in Washington DC last month. Holt said that the Democratic leadership had asked for data showing the impact the research spending would have on jobs before agreeing to up the funding for science agencies in the bill. Such data, he said, were not readily available, although after some scrambling, agencies were able to cobble together rough figures sufficient to carry the day.

But the data question is not about to go away. Indeed, the stimulus bill (now the Recovery Act) means that gathering data on the short-term impact of research spending on jobs is about to become a preoccupation of the federal science agencies and their beneficiaries. Under the act, each grant recipient is required to report to the government quarterly on the number of jobs created and the number retained as a result of the stimulus money. The White House Office of Management and Budget is developing guidelines that will govern exactly how this information will be calculated, gathered and made public. But it's already clear that the reporting will probably go beyond existing efforts, in which agencies collect information, at most, about how many individuals were supported by a grant.

There's an old saying that what you measure is what you get. So, will this focus on near-term job creation change the way science agencies go about their business or how they're evaluated? It could. At a recent hearing of the House Committee on Science and Technology, Congresswoman Kathy Dahlkemper, a first-term Democrat from an economically hard-hit section of Pennsylvania, asked whether science agencies were "taking into consideration what areas of the country have the greatest need for job creation" when awarding stimulus funds. This would be a poor way to distribute the stimulus money, but it's not a ridiculous question to ask about a law that was explicitly presented as a way to create jobs now.

And the agencies' answers showed how much the rhetoric around the law is shaping their actions and how much fodder could end up being provided for future debates. Cora Marrett of the National Science Foundation (NSF) said her agency had mapped out where proposals already in hand that could be considered under the Recovery Act had come from, and that the NSF wanted to be sure it was "addressing needs, as those might vary across the country". Matthew Rogers of the Department of Energy said that "every dollar under the Recovery Act is associated with" a specific number of jobs, a state and an impact, adding that job creation and retention would be tracked by congressional district.

Concern about the geographical distribution of research funding is not new; whether federal dollars would flow almost exclusively to old-line Northeastern universities was a subject of debate when Congress created the NSF in 1950. And complaints about the geographical concentration of federal science money have been among the justifications for congressional earmarks — money Congress directs to specific projects at locations or institutions it selects. Indeed, the number of academic earmarks has skyrocketed in part because of a previous rhetorical gambit. In the 1980s, the scientific community began describing universities as economic development tools because money was being handed out to spur 'competitiveness'. But associating individual grants with specific metrics about employment and considering near-term job creation as a rationale, or even a criterion, for making awards takes this line of thinking further than it's ever gone before. And all the information on jobs will be readily available on the web, displayed to a public that is in a sceptical and populist mood in the wake of bonus payouts to financial companies.

The Recovery Act is also prompting efforts to develop more rigorous economic analysis of the impact of science spending on jobs. The NSF's Science of Science and Innovation Policy Program has issued a call for proposals for research to evaluate the impact of the stimulus bill including such questions as: "What was the contribution of the science investment to the creation and retention of jobs?" (In the worst-case scenario, the lack of data about job creation will be replaced by clashing economic theories about it.)

Immediate job creation will not be the sole measure of the success or failure of the stimulus package, although it will no doubt be the most politically salient metric. But even broader means of evaluating the Recovery Act have a short-term focus because of the way the bill was sold. At the science committee hearing, Brad Miller, the North Carolina Democrat who chaired the session, said that "when the stimulus funds run out next year", Congress will want to know "did they provide investments needed to increase economic efficiency, by spurring technological advances in science and health". That may not be easy to know after just two years. The NSF research programme is also interested in proposals exploring "what scientific or technological advances" were achieved with the stimulus funds, but doesn't necessarily expect such advances to show up immediately.

It's obviously too soon to know whether the stimulus experience will change the way science funding is viewed in a significant or lasting way. And the pressures will vary depending on each agency's mission. But it is soon enough to conclude that the stimulus debate has underscored the importance of an oft-forgotten lesson in Washington: when you come up with a line of argument, think about what would happen if people actually believed you. ■

David Goldston is a visiting lecturer at Harvard University's Center for the Environment. Reach him at partyofonecolumn@gmail.com

See also page 556.

NATURE|Vol 458|2 April 2009

ONE HUNDRED YEARS OF RITA

From a home lab to the Italian Senate, by way of nerve growth factor — Rita Levi-Montalcini is a scientist like no other. **Alison Abbott** meets the first Nobel prizewinner set to reach her hundredth birthday.

Tiny though she is, Rita Levi-Montalcini tends to command attention. And on the morning of 18 November 2006, she had the attention of the entire Italian government. A senator for life, Levi-Montalcini held the deciding vote on a budget backed by the government of Romano Prodi, which held a parliamentary majority of just one.

A few days earlier, Levi-Montalcini had said she would withdraw her support for the budget unless the government reversed a last-minute decision to sacrifice science funds. It was Levi-Montalcini versus Prodi — and Levi-Montalcini won. On the morning of the vote, immaculately turned out as always, she walked regally on the arm of an usher to her seat in the Italian senate and cast her vote. At one stroke, she secured the budget, won a battle for Italian science and snubbed Francesco Storace, leader of the Right party and part of the opposition coalition. A few weeks earlier, Storace had caused a national scandal by announcing his intention to send crutches to Levi-Montalcini's home — symbolic of her both being a crutch to an ailing government, he said, and her age, which he considered too old to be allowed to vote.

Levi-Montalcini didn't consider herself too old then, when she was 97 years old, and she certainly doesn't now when, on 22 April, she will become the first Nobel laureate to reach the age of 100. Italy — and quite possibly the world — has never seen a scientist quite like her.

Born into a well-to-do Jewish family in Turin in 1909, Levi-Montalcini fought hard for her career from the beginning. First there was her domineering father, who didn't believe in higher education for women. Then there were Benito Mussolini's race laws, which ejected Jews from universities and forced her into hiding. And after that there was the scientific establishment, which refused to believe in the existence of nerve growth factor (NGF), the discovery of which eventually won Levi-Montalcini a share of the 1986 Nobel Prize in Physiology or Medicine, together with her colleague Stanley Cohen. "That discovery was huge — it opened up a whole field in understanding how cells talk and listen to each other," says neuroscientist Bill Mobley of Stanford University in California, an admirer for more than 30 years. Hundreds of growth factors are now known to exist and they affect almost all facets of biology.

Despite her age, Levi-Montalcini still works every day, exquisitely dressed, hair stylishly coiffured, hands perfectly manicured. In the mornings she shows up at her namesake European Brain Research Institute (EBRI)–Rita Levi-Montalcini, on the outskirts of Rome. In the afternoons she goes downtown to the offices of an educational foundation for African women that she created in 1992.

Turning 100 is no reason to stop fighting. "It's not enough what I did in the past — there is also the future," Levi-Montalcini says. She has never hesitated to use her Senate position to push for better scientific prospects in the country. And today she has something even closer to her heart to fight for — the survival of the EBRI, which she created in 2002 and which is now in financial straits.

Levi-Montalcini spent a large part of her research career in the United States. But her early, and late, scientific life has been based in Italy. Three years after leaving high school, she finally persuaded her father to allow her to study medicine, and in 1930 she enrolled at the University of Turin. Her first mentor was Giuseppe Levi, a prominent neurohistologist. In her autobiography *In Praise of Imperfection*, Levi-Montalcini refers to him as "the Master" — he was an outspoken antifascist, renowned for his alarming fits of rage. But he was also the man who introduced her to her first passion: the developing nervous system. Under Levi's attentive eye, she mastered a technique that would be key to her own successes, that of silver-staining nerve cells. Developed by Camillo Golgi in the late nineteenth century and later refined by the Spanish neuroscientist Santiago Ramón y Cajal, the technique allowed individual nerves to be seen under the microscope with perfect clarity.

SPL

Growing up, Levi-Montalcini fought her father to be able to attend medical school.

Levi-Montalcini's independent research started when Mussolini's race laws were passed in 1938, and all Jews were expelled from universities and other public institutions (Levi, too, was thrown out). Inspired by the story of Cajal, who had worked alone in a makeshift lab in out-of-the-way Valencia, she set up a bedroom laboratory at her family home. When Levi returned to Turin some time later, he joined her at her bedroom bench.

She had already identified her research challenge: to work out how nerves emerging from the embryo's developing spinal cord find their way to the budding limbs they will eventually innervate. She had recently come across an exciting paper[1] published a few years earlier by embryologist Viktor Hamburger at Washington University in St Louis, Missouri. Hamburger had removed the growing limbs of chick embryos and found that doing so reduced the size of the ganglia, tiny structures that cluster together the nerve fibres emerging from the spinal cord and direct them on to their final destinations. He put this atrophy down to the absence of what he called an inductive factor released by the tissue to be innervated and, he proposed, necessary to make precursor cells proliferate and then differentiate into neurons.

Detailed dissections

Hamburger, though, could not see the nerve fibres in great detail using the light microscope. So Levi-Montalcini decided to repeat the experiment with the silver-staining method. Like Cajal, she reasoned she would need little more than an incubator and a microscope — and a regular supply of fertilized hen's eggs. Using tiny scalpels and spatulas fashioned out of sewing needles to do her dissections, she saw that the ganglia did not, in fact, wither immediately. The neurons actually proliferated, differentiated and started to grow towards their targets. It was just that they died before reaching them. She concluded that the problem was not the lack of an inductive factor, but of a growth-promoting one that would normally be released by the budding limbs[2].

Towards the end of 1942, bombing forced the Levi-Montalcini family to move into the countryside, where she continued her research undaunted, cycling to farms to buy fertilized eggs. She stopped only when Italy switched allegiance to the Allies in 1943, and Hitler's troops invaded northern Italy.

After the war, Levi-Montalcini returned to Turin as Levi's assistant. But at 36, the role no longer suited her — after all, he had been an occasional assistant to her in the days of her bedroom lab. She found her way out when Hamburger, who had read the papers she had published with Levi during the war, invited her to St Louis for a semester to repeat and extend her experiments.

Just as she was doing those experiments, something happened that extended her stay in St Louis from one semester to 26 years. One of Hamburger's graduate students, Elmer Bueker, was trying to see if any piece of fast-growing tissue could attract nerve fibres in the same way that fast-growing developing limbs do. He grafted a lump of proliferating mouse sarcoma

ALBERT WATSON



BECKER MEDICAL LIBRARY, WASHINGTON UNIV. SCHOOL OF MEDICINE

ANSA/ALINARI

Levi-Montalcini worked at Washington University through the 1950s (left) and 1960s (middle). In 1986, she and Stanley Cohen (seated either side of table) were awarded a Nobel prize .

tumour onto a chick embryo and found that nerve fibres grew and invaded the tumour mass more abundantly than the limb bud. He postulated that the greater surface area of the tumour allowed more nerves to grow up to it.

Levi-Montalcini is renowned for her exceptional intuition, and Bueker's experiment made her antennae vibrate. To her eye, the invasion did not look quite right. Although nerves grow into developing limbs in an orderly way, their growth into the tumour was massive and wild, with the fibres branching randomly. She became convinced that the transplanted tumour tissue was releasing the same sort of factor she claimed the developing limbs released, a factor able to diffuse to the ganglia and stimulate the growth of nerve fibres.

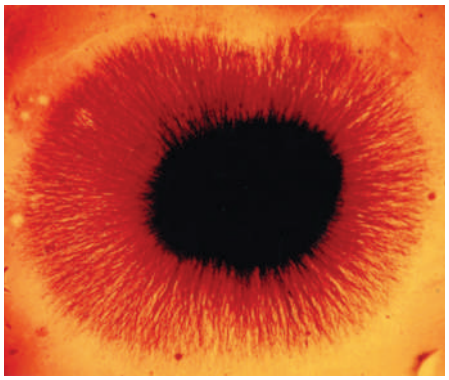
Inspired insight

She repeated the experiment, ingeniously placing the tumour outside the sac containing the embryo. This area, although physically separate, shares the embryo's blood supply. It was a killer experiment. Nerves sprouted and grew wildly, supporting her theory that the tumour was releasing a factor that diffused into the blood and travelled to the embryo[3]. "She realized there was another way to interpret the data, and she knew what had to be done," says Lloyd Greene, who studies neuronal differentiation at Columbia University in New York, and has known Levi-Montalcini since he was a student.

But to really prove her point, Levi-Montalcini needed a system that was more reliable and flexible than the fertilized egg, and one that would allow her to quantify the responses she was measuring. She wanted to learn how to culture isolated chick-embryo ganglia, and knew of only one laboratory that could do so. So she put two live, tumour-riddled white mice into her handbag and boarded a plane for Rio de Janeiro, where another of Levi's former students was running a big tissue-culture facility.

In Rio she learned to culture isolated ganglia and she grew them close to pieces of mouse sarcoma. After 24 hours of culture, she was thrilled to see haloes of nerve fibres growing from the ganglia like suns, with their highest density facing the tumour. Her many letters to Hamburger include beautiful drawings of the haloes. Levi-Montalcini's strong artistic bent is also evident in her research papers, which she illustrated by hand, and in the clothes that she designs for herself.

By the time she returned from Rio, Cohen had joined the Hamburger group. The pair worked together for six years trying to identify



Halo effect: Levi-Montalcini found that a growth factor causes nerves to sprout from chick ganglia.

the factor released by the tumour. Both were determined to provide the sceptical scientific community with solid chemical evidence that the nerve-promoting factor was a reality. But scepticism only increased when Cohen and Levi-Montalcini proposed that snake venom and extracts of mouse salivary glands, both of which also promoted profuse nerve growth, were abundant sources of the factor they were seeking.

For many scientists, it required too great a leap of the imagination to believe in this unlikely soluble factor, which was supposed to diffuse from one tissue and then potently affect specific processes in nerves. "You have to remember that such a mode of biological action was not accepted in those days," recalls Ralph Bradshaw, who joined Washington University in 1969 as its first protein chemist and is now at the University of California, San Francisco. "And Rita was saying it was in tumours, snake venom, as well as many normal tissues — well, people just didn't believe it was serious biology."

More people started to believe when Cohen discovered another, related factor that was later called epidermal growth factor[4]. Then, in 1959, Levi-Montalcini developed with him an antiserum to purified NGF. The antiserum abolished the *in vitro* halo, and wiped out the relevant part of the nervous system when injected into newborn mice[5]. The last remaining pockets of scepticism in the scientific community dissolved when Bradshaw, together with Ruth Hogue Angeletti, the only PhD student Levi-Montalcini ever had, determined the structure of the protein in 1971 using one of the first automated protein sequencers[6]. "Rita didn't put her name on the paper as we would have expected someone in her position to do," says Bradshaw. "A typical Rita gesture."

Although Levi-Montalcini loved the scientific atmosphere in the United States, she was always homesick for Italy and for her family. In the early 1960s, she began to split her time between St Louis and Rome, where the CNR, Italy's major research organization, created a laboratory for her. Her working style was relentless, demanding and passionate. In the decades in which her research was most intense, she would call her co-workers before seven in the morning as well as last thing at night to discuss experiments. Angeletti refers to the regime as inspiring rather than brutal. "Even as a highly motivated young American I had never before observed this kind of dedication," she says. "I realize how lucky I was to work with someone so brilliant, expansive and generous of spirit."

When the study of growth factors finally became respectable and other scientists flooded into the area, rather than being gratified, Levi-Montalcini was annoyed by the invasion of what she saw as her territory. "She fell out with most

R. LEVI-MONTALCINI

C. CABROL/KIPA/CORBIS

M. SIRAGUSA/CONTRASTO/EYEVINE

Levi-Montalcini has published 21 popular books and continues to work at her namesake brain research institute in Italy (right).

people in the NGF field at one time or another — including myself," recalls Bradshaw. At meetings, she had a tendency to educate audiences on the order in which discoveries had been made, recalls Greene. After one of his own talks, hers was the first hand raised. "It was not a question, but a long statement about NGF and its history," he says. "As she spoke, she little-by-little made her way to the stage and the podium, and the next thing I knew, she was next to me at the microphone still asking her 'question'." Under the circumstances, he says, he could do no more than "step aside, cede the microphone to her, raise my eyebrows and let her finish".

Peacemaker

In the early 1980s, Levi-Montalcini started to bury the hatchet with everyone in the field, says Bradshaw. Their own quarrel — over a paper he had published without showing her first — was patched up when she took him aside for a chat at a meeting. "It ended what had been a strained and difficult time for me," says Bradshaw. "But Rita had to endure a great deal of scepticism in the early days and there were times when she was justifiably defensive." Her later discoveries faced no such scepticism. She showed, for example, that NGF had major effects on the immune system, yet another unexpected finding that became a major turning point in biology[7].

By the time she and Cohen were awarded the Nobel prize, considerable peace had been achieved. But controversy picked up again in the wake of the award. Some were upset by what they saw as her failure to acknowledge her debt to others, such as Levi and Hamburger. Hamburger, who lived to be 100, claimed that their friendship suffered after she explained publicly why he should not have shared the prize with her as some had thought appropriate.

But such criticism gained no traction in Italy, where Levi-Montalcini had by now settled permanently. Many viewed her as a national treasure for her achievements, outsize personality, energy and eloquence. Her CNR institute became one of the largest biological research centres in the country. She also took it on herself to work at all levels to improve the state of Italian science. A socialist by lifelong conviction, she became good friends with Prodi, who had been prime minister in two centre-left governments. After she was made senator for life in 2001, she showed up for every parliamentary vote to support Prodi's fragile coalitions.

She also champions social issues related to research, such as ethics and women in science. The Rita Levi Montalcini Foundation has supported education for more than 6,000 African women — "to improve their chances of becoming scientists", she says. A keen writer, she has published 21 popular books. As a young bookworm, her favourite among the classics was Emily Brontë's tale of dark passion, *Wuthering Heights*. Such romantic inclinations remained literary though — despite a brief engagement while at medical school, she never had any long-term romances. In a 1988 interview with *Omni* magazine she said, tellingly, that even in a marriage of two brilliant people, "one might resent the other being more successful".

"If I die tomorrow or in a year, it is the same — it is the message you leave behind you that counts."
— Rita Levi-Montalcini

One of her remaining desires has been to leave as a legacy a well-run research institute of international significance in her country, where underfunding, inefficiency and bureaucracy have crippled much of the state research system. The Santa Lucia Institute in Rome, keen to expand its own research activities, offered rent-free premises for the first ten years of her neuroscience institute. But the EBRI is now looking shaky. Levi-Montalcini expected the government to make funds available for running the institute, but in the event the Prodi government provided only a one-off donation of €3 million (US$4 million) just before its demise one year ago — and no other major donor was found. The right-wing government of Silvio Berlusconi has shown little interest in research and the name Levi-Montalcini cuts no ice with it.

The EBRI, which now has a staff of 28, runs with an annual deficit of €200,000. Earlier this year, University of Turin neuroscientist Piergiorgio Strata took over as scientific director with a mandate to turn things around. "We need maybe €3 million per year to survive," says Strata, who is confident that he'll be successful. The ever-determined Levi-Montalcini puts her trust in him. "I'm an optimist," she says. "I still hope we can find a way to carry on."

Levi-Montalcini is now hard of hearing and sees poorly, but her mind is sharp. At the EBRI she runs a research project to see how far back NGF goes in evolution. Several young scientists are helping by trying to find out whether the factor exists in a series of invertebrates. They are gratified to be able to speak with her most days. "She is an inspiration for us," says Francesca Paoletti, one of the postdocs working there.

And they, in turn, make her happy. "I am not afraid of death — I am privileged to have been able to work for so long," says Levi-Montalcini. "If I die tomorrow or in a year, it is the same — it is the message you leave behind you that counts, and the young scientists who carry on your work." And with that, clutching her micrographs of NGF in octopus tissue, she walks away on the arm of a friend, with a slow but stately gait. With her high heels and the swing of her tailored coat, she still looks as though she stepped off the pages of a fashion magazine. ■

Alison Abbott is *Nature*'s senior European correspondent.

1. Hamburger, V. *J. Exper. Zool.* **68,** 449–494 (1934).
2. Levi-Montalcini, R. & Levi, G. *Arch. Biol. Liège* **54,** 189–200 (1943).
3. Levi-Montalcini, R. *Ann. N. Y. Acad. Sci.* **55,** 330–343 (1952).
4. Cohen, S. *J. Biol. Chem.* **237,** 1555–1562 (1962).
5. Levi-Montalcini, R. & Booker, B. *Proc. Natl Acad. Sci. USA* **46,** 384–391 (1960).
6. Angeletti, R. H. & Bradshaw, R. A. *Proc. Natl Acad. Sci. USA* **68,** 2417–2420 (1971).
7. Levi-Montalcini, R. *et al. Progr. Neuroendocrinol.* **3,** 1–10 (1990).


A. MARTIN

The textbook of the future

Undergraduate textbooks are going digital. **Declan Butler** asks how this will shake up student reading habits and the multi-billion-dollar print textbook market.

The rumble of textbooks thumping on to the desks of a university lecture theatre, the rustle of turning pages, the groan of backpack straps hoisting 10 kilograms of textbooks — these sounds may soon be an echo of the past. This semester, 1,200 students at the University of Texas at Austin (UTA) are foregoing printed textbooks in a pilot trial of Amazon Kindle e-readers stuffed with texts in electronic form. At NorthWest Missouri State University (NWMSU) in Maryville, classes are testing textbooks on Sony e-readers, as well as on the students' own laptops, as part of plans to roll out e-textbooks across all courses within 5 years. The list goes on: within the past 18 months or so, as textbook publishers have begun to make more and more titles available online, universities worldwide have begun to experiment with e-textbooks.

"E-textbooks are not yet mainstream — but they are on the edge of a breakthrough into the mainstream," says Kevin Hegarty, UTA chief financial officer. Indeed, textbook publishers are scrambling to position themselves for a revolution in the way they do business as they rethink their decades-old model of massive, printed tomes sold at premium prices.

The resulting proliferation of new models — none of which is yet a sure winner — is being shaped by the interplay of at least three forces: new e-readers and displays for viewing and interacting with the e-textbook content; new business and licensing models for delivering quality content at prices students and universities can afford; and new concepts for the content itself, and for how it is created.

Beyond black and white

On the hardware front, e-textbooks are reaping the benefits of rapid innovation in electronic readers for documents and novels. Most of the latest generation of e-readers, such as Amazon's Kindle 2 and Sony's PRS-700, offer displays based on technology from the E-Ink Corporation of Cambridge, Massachusetts (see *Nature* doi:10.1038/news.2009.202; 2009). These displays produce text and images that rival the brightness and clarity of ink on paper, which makes reading them far more comfortable than reading text on the liquid crystal display screens of laptops and desktop computers. They also allow an e-reader's batteries to last for days: the displays require power only when the screen is being changed — for example, by 'turning' a page. The first generation of such e-readers, launched less than three years ago, has already sparked mass uptake of e-books, and they could potentially do the same for e-textbooks.

As delivery vehicles for textbooks, however, existing e-readers still leave a lot to be desired. For example, most are designed for reading books from beginning to end. But "very few students read a textbook in that manner", says Paul Klute, who is directing the NWMSU e-textbook project. He recalls how the school launched its pilot test of the Sony's PRS-505 reader in autumn 2008 with e-textbooks from six publishers. It was an instant flop with the 200 student testers. They wanted to do what they had always done, says Klute, and flip through to find bits they didn't grasp in the lecture, or dip in to read short sections, or find a key figure. But the e-reader wasn't built for this, so they ended up frustrated. This semester,

Sony has replaced the device with the newer PRS-700. Its search and navigation functions and the ability to flip a page by swiping a finger across the touch screen have elicited a much more positive response, Klute says.

Another drawback of current e-readers is that they have small black-and-white displays, just a little larger than 9 by 12 centimetres. This makes them unsuited to most science textbooks, which typically have large pages and colourful graphics. "The market is not likely to expand until the e-readers improve," says Hegarty.

Many large textbook companies are holding off from experimenting with e-readers until that happens. But manufacturers promise that big screen, colour e-readers are on the way within a year or two. If so, this will be the tipping point at which e-textbooks take off, predicts Hegarty. "It will be a big leap forwards," he says.

If the price is right. Dedicated e-readers currently start at prices of around US$350, points out Joe Esposito, a digital-media consultant and former chief executive of *Encyclopaedia Britannica* online. Reading an e-textbook on a laptop might not be as easy on the eyes, but most students already own a laptop — complete with a colour display. "The student laptop will prove a potent competitive entry barrier to other devices for reading e-textbooks," says Esposito. This is why NWMSU is also piloting e-textbooks on laptops among 500 students in 11 disciplines in an effort to compare how well students learn with e-readers, laptops and print textbooks.

That is probably a wise approach. Five years ago, devices such as the Kindle did not even exist. Which devices students will use for reading e-textbooks five years from now is anybody's guess — although many people are betting on some sort of convergent evolution among e-readers, laptops, portable music players and smart phones. The boundaries will increasingly blur, predicts Neelan Choksi, co-founder and chief operating officer of Lexcycle, a company based in Portland, Oregon, that makes Stanza, a popular e-book reader application for the iPhone. "Everyone is racing to be the ultimate multi-function device," he says.

Kindling a revolution

But device innovation has other implications as well. Just as the Internet brought dramatic change to the music industry, which relied on selling content on a physical medium, such as the CD, better devices could similarly disrupt the textbook industry. So it is not surprising that textbook publishers' embrace of e-textbooks is reminiscent of two scorpions mating.

Like the music industry, textbook publishers have been reluctant to put content online because of concerns about piracy, and the risk that it might undermine sales of their traditional print editions. If they are now willing to do so, it is largely because such concerns have been offset by the realization that e-textbooks may give them a way to cut into the largest threat to their profits: the huge market for second-hand textbooks.

> **"Everyone is rushing to be the ultimate multi-functioning device."**
> — Neelan Choksi

Thanks to the Internet, what was once the preserve of local used bookstores is now a vast and sophisticated international online market. The US market for new textbooks is estimated at around $5.5 billion, but the parallel market for used books is around one-third of that, says Esposito. Publishers hope that by offering lower priced e-textbooks they can obliterate the used-textbook market, from which they currently get nothing, and sell electronic versions semester after semester — presumably with frequent updates, analogous to the new print editions they regularly bring out.

J. LEE/BLOOMBERG NEWS/LANDOV/PA

Amazon's Kindle: bringing technology to book.

But publishers' enthusiasm for e-textbooks remains relative, says Esposito. "E-textbooks are too big a market for publishers to walk away from, but publishers are not willing to walk away from the print market that makes up more than 90% of their sales." This defence of the print market is reflected in their offerings, which are usually electronic facsimiles of printed textbooks, sold to students online, and which provide only the most basic functionality, such as printing, highlighting and making electronic annotations.

By far the largest market for textbooks is the United States, and the companies that win in this space are also likely to be those that will dominate worldwide. Because of this, it is also likely to be where the evolution of e-textbook business models plays out.

The biggest player is CourseSmart, a consortium in Belmont, California, created by the five publishers who together account for roughly 85% of the global print textbook market: Pearson; Cengage Learning; McGraw-Hill Education; John Wiley & Sons; and the Bedford, Freeman & Worth Publishing Group. (The last is a unit of Macmillan, which is owned by *Nature*'s parent company, the Georg von Holtzbrinck Publishing Group based in Stuttgart, Germany.) "We have brought a critical mass of textbooks together on a single common platform for the first time," says Sean Devine, chief executive of CourseSmart.

CourseSmart sells its e-textbooks at about half the price of its print versions, and so far has made more than 5,800 e-textbooks available at its website, or about one-third of the world's

PLASTIC LOGIC

Students say they would prefer to have print textbooks — until they are offered a cheaper option.

most popular textbooks. Students who buy the books are constrained by digital rights management. The copy they buy usually 'expires' after their course has ended, after which it no longer accessible. CourseSmart's digital rights management also forbids students from moving a book downloaded on one computer to another device, limits printing to 10 pages at a time, and allows the whole book to be printed only once.

Bulk buying

Nonetheless, student purchases of CourseSmart e-textbooks are growing rapidly, says Devine. A survey by NWMSU in February found that, all things being equal, about half the students would prefer print textbooks and about a quarter would prefer e-textbooks, whereas the remainder had no strong feeling. But when asked what they would do if buying a textbook themselves, almost 80% said they would opt for the cheaper e-textbook offering.

Ongoing tests of CourseSmart e-textbooks by the University System of Ohio show that they reduce costs — the average US student forks out some $900 annually on print textbooks — and students using them perform just as well as when using paper versions, says Peter Murray, deputy head of new service development at the Ohio Library and Information Network in Columbus, Ohio, which assists the University System of Ohio on the project.

But Make Textbooks Affordable, a coalition of US student groups, thinks that students are being fleeced, and that the price of 'renting' an electronic file, which costs little for publishers to distribute, is excessive. Indeed, if an e-textbook typically costs half that of the print version, the saving is less impressive when one considers that buyers of new print books would recoup much the same by reselling, and students might pick up used versions for the same price or less.

Charging half the price of a printed textbook for an e-book that expires is "far too costly", says Hegarty. Rather than leaving students to act as isolated agents in the marketplace, he says, universities, or consortia of universities, should step in and use their bulk-purchasing clout to force down prices by negotiating site licences to e-textbooks, just as many do for online versions of scientific journals. E-textbooks procured this way could be made free at the point of use to all on campus, or for flat fees included in tuition fees. "The winning model will involve licensing content broadly such that the library licenses the materials, the professors assigns them and the student electronically checks them out of the library as they do hardcopy books," he says.

The multi-function iPhone: one ring to rule them all?

M. LENNIHAN/AP

Klute also favours such a scheme. NWMSU already spends around $800,000 a year on tens of thousands of copies of print textbooks that it rents to students, who are charged $80–$90 per semester for textbook provision. He thinks that using an e-textbook site licence could at least halve that cost to students.

Such a model is being tested by the UK National E-books Observatory project. The project has licensed from publishers 36 e-textbooks in business and management, medicine, media studies and engineering from September 2007 to August 2009 at a cost of £600,000, and made them available free to all UK universities. It is the future, says Liam Earney, collections team manager of the Joint Information Systems Committee, based in London — a body established by Britain's higher-education funding councils to support education by promoting technological innovation — which operates the pilot.

> **"Cheap prices are the most effective digital-rights management."**
> — Eric Frank

Open source

A more radical idea is to offer textbooks for free, without rights restrictions. A range of free, open textbooks are already available for download at WikiBooks (http://en.wikibooks.org); the Community College Consortium for Open Educational Resources' Open Text-Books Project; and Connexions, created in 1999 by electrical engineer Richard Baraniuk of Rice University in Houston Texas. These texts typically take the form of modules written by many expert authors.

For now these free textbooks remain a cottage industry, says Esposito. Wikipedia-like volunteer efforts are much better suited to self-contained modules that are small enough for an individual to see through from A to Z. But a textbook demands a coherent overall structure and coordination between sections. That is why creating one has always been a major undertaking, demanding long-term commitments by publishers — who need to make a profit — and by authors who usually want to be paid for their effort.

Still, perhaps 'free' and 'profitable' need not be a contradiction in terms. One group of veteran textbook publishing executives is trying to put open textbooks on a solid commercial footing. In 2007 they created Flat World Knowledge, based in Nyack, New York, and in January 2009 rolled out the first of the 21 textbooks they have in development so far. The texts are written by some 40 domain experts who will be paid 20% of royalties. The company also plans to make its content available via Kindle and other e-readers. All its content will be free to reuse for non-commercial purposes under a creative commons licence.

Eric Frank, Flat World's co-founder, says that the strategy is to attract greater use by giving the e-textbooks away — the initial targets are the high-volume texts for first-year students — and then look for profit from students' purchase of print-on-demand versions at $29.95 for black and white, and $59.95 for colour. Students can copy and use the electronic content in any way they wish, says Frank. "Cheap prices are the most effective digital-rights management," he says. "We want to avoid a digital-rights war with students." The company also hopes to make money by licensing its content to commercial companies, such as distance-learning outfits and course-management software firms.

By making its content free for reuse, Flat World Knowledge will allow lecturers to splice and dice its content. "More and more professors want to teach from 'customized' textbooks, which are aggregations of various materials, not just what a publisher has aggregated in a single book," says Hegarty. He says that the UTA has made an electronic tool available for academics to aggregate any licensed library materials, including scientific journals, and 'publish' them to their students as their textbook materials. "I think that this is where textbooks are headed."

In the larger sense, of course, no one really knows where e-textbooks are headed. They just know that things are moving very fast. About all that's certain, says Klute, is that the next chapter of e-textbooks is now being written. "E-textbooks as we currently know them will look drastically different five years from now". ■

Declan Butler is a senior reporter at *Nature*, based in France.

See Editorial, page 549.

CORRESPONDENCE

Austria should invest in brains, not in bricks, banks or airlines

SIR — Because of uncertainty about this year's science budget, as expressed in your News in Brief story 'Austrian scientists rattled by threat to funding' (*Nature* **457,** 648; 2009), the Austrian science fund FWF has postponed its first two board meetings of 2009. It has frozen all decisions on already-reviewed grant applications until May 2009. As the FWF is by far the most significant public agency supporting basic research in Austria, any reduction of its moderate budget would be a devastating blow.

This uncertainty puts the Austrian government's recent efforts to advance science, and to attract internationally renowned scientists, into serious jeopardy. Because the basic subsidy for universities is low, scientists have been relying heavily on competitive funding from the FWF.

We find it obscene that the government is pursuing its plan to establish an 'elite university' near Vienna — the Institute of Science and Technology Austria — while competitive funding is at risk. Do the institute's newly appointed president and his senior academic staff know that one crucial pillar of their budget is cracking? Do our gifted young students preparing for an academic career recognize that, without funding by the FWF, the academic world is at risk? Do their parents know that their children are heading down a blind alley?

We hope that our officials consider what is best for the future of the country: invest in brains — not bricks, banks or airlines. Knowing what technology means for a country, the US National Institutes of Health has just received additional funding of $10.4 billion. Perhaps Austrian students and scientists will again have to go west to the United States to survive the current global economic crisis.

Michael Freissmuth Department of Pharmacology, Medical University of Vienna, Währingerstrasse 13A, 1090 Vienna, Austria
e-mail: michael.freissmuth@meduniwien.ac.at
Sigismund Huck Center for Brain Research, Medical University of Vienna, Spitalgasse 4, 1090 Vienna, Austria

Evolution and intelligent design in Hong Kong

SIR — Your News story 'Hong Kong evolution curriculum row' (*Nature* **457,** 1067; 2009) reports a call by faculty members at Hong Kong University for a sentence to be removed from new guidelines for secondary-school biology education. At present, these state: "In addition to Darwin's theory, students are encouraged to explore other explanations for evolution and the origins of life, to help illustrate the dynamic nature of scientific knowledge". You also note that a professor criticized the university for not letting him teach intelligent design in his course on the origin of the Universe.

I was born in Hong Kong, was educated at local missionary schools and Hong Kong University, and am now overseas doing research in the field of evolution and development. As a scientist, I believe that the purpose of education is not only to pass knowledge to future generations, but also to develop students' analytical and critical thinking. Central to both aspects is the need to focus on facts and testable views supported by evidence. This is all the more important given the limited amount of time available for teaching and its support by public funding. Evolution fulfils these necessary criteria, whereas intelligent design, being untestable and unsupported by evidence, does not.

Hong Kong is a multicultural society, deeply imprinted with traditional Chinese culture and values, but also facing a constant inrush of ideas from the West. The fundamental cause of these controversies is more than just a cultural clash. It reflects a lack of long-term public education in evolutionary biology. In the year of Darwin 200, it is time to rectify this situation.

Jerome H. L. Hui Faculty of Life Sciences, University of Manchester, Michael Smith Building, Manchester M13 9PT, UK
e-mail: jerome.hui@manchester.ac.uk

Scientists must stand up and be counted

SIR — In your Editorial 'Against vicious activism' (*Nature* **457,** 636; 2009), you call for scientists and the authorities to stand up for animal research in basic and applied science. However, you may be putting the cart before the horse in recommending that officials and politicians become advocates of animal research in order to encourage individual scientists to do so.

In the United Kingdom, it was the actions of individual scientists — and of members of the public who joined the Pro-Test demonstration in Oxford in February 2006 and signed the Coalition for Medical Progress's petition — that gave politicians and other public figures the encouragement they needed to come out in support of animal research. The lesson to be learned from the UK experience is that scientists at the universities being targeted by extremists, alongside students and advocacy groups, must be encouraged to stand up and be counted. Only then can they expect others less directly involved to take an unequivocal public stand.

A parallel could be drawn with the debate over the use of embryonic stem cells for research in the United States, where support among the general public and in Congress has been driven by the strong vocal endorsement of individual scientists and advocacy groups.

The truth, uncomfortable though it may be, is that — as with many controversial areas of science — those working with animals in research must make a public case to justify their use, and must be willing to show unequivocal support for colleagues who speak up. Do that, and the rest will follow.

Dave Bienus Speaking of Research, and Pennsylvania State University, 101 Centralized Biological Laboratory, University Park, Pennsylvania 16802, USA
e-mail: dab43@psu.edu

Animal-health facility in Germany leads the way for Europe

SIR — Your News story 'Britain hits a hurdle in replacing key animal-pathogen facility' (*Nature* **457,** 769; 2009) describes the problems faced by the Institute for Animal Health in Pirbright. It is deplorable that this world-class institute is uncertain of being able to develop a key animal-pathogen facility and other adequate infrastructure.

In contrast, Germany's federal ministry of food, agriculture and consumer protection is investing nearly €300 million (US$395 million) to create a state-of-the-art facility for infectious-disease research at our institute on the Isle of Riems in the Baltic Sea. New laboratory and animal facilities will be constructed, including a biosafety-level-4 facility for large animals that is unique in Europe. The plans were developed in the mid-1990s and construction should be largely finished in time for the institute's centenary in late 2010.

Thomas C. Mettenleiter Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Südufer 10, 17493 Greifswald-Insel Riems, Germany
e-mail: thomas.mettenleiter@fli.bund.de

ESSAY

All the President's scholarly men

Barack Obama's choice of science advisers is cause for celebration. Yet history shows that an impressive academic record doesn't guarantee good, impartial advice, cautions **Robert Dallek**.

President Barack Obama's appointment of academic scientists and economists to positions of high authority in his administration has created the sort of excitement in universities and among researchers that has not been seen for eight years. Certainly, after George W. Bush's grudging agreement to a constricted programme of stem-cell research and his politicization of scientific findings about the environment, Obama's choice of prominent scholars is a breath of fresh air.

Yet before the country's, or indeed the world's, academics become too excited about the latest professors at the White House, they would do well to recall that US presidents have repeatedly turned to academic stars for advice during the past century, with mixed results. That academics have an imperfect record as presidential advisers is not to doubt that their expertise has considerable value. But no one should assume that an impressive academic track record guarantees good policy. Far more important is an ability to remain independent and offer advice based on sound evidence.

The good and the bad

Among the most striking achievements by academic insiders in presidential administrations is that of the 'Brain Trust', a group of Columbia University professors who counselled Franklin D. Roosevelt on how to repair the damage caused by the Great Depression[1]. Also high on the list in importance is the work of Robert Oppenheimer, a physicist from the University of California, Berkeley.

In June 1941, almost two years after Albert Einstein had alerted President Roosevelt to the possibility of building an atomic weapon, Roosevelt created an Office of Scientific Research and Development. Oppenheimer became the chairman there of a subcommittee charged with designing the A-bomb. In March 1943, when the US Army selected Los Alamos, New Mexico, as the site where the work would be done, Oppenheimer became the principal architect of the weapon.

Fear that Hitler's Germany would claim victory in the race for the 'winning weapon', as some called it, were overblown, although the limits of Germany's capacity to build a bomb were not fully understood until later. Despite this — and retrospective qualms voiced by Oppenheimer and several of his colleagues about building so destructive a weapon — the United States' success in designing, testing and using the A-bomb was testimony to an extraordinary cooperation between the federal government and the scientific community[2].

A comparable success story is Henry Kissinger's role in shaping some of Richard Nixon's foreign policies. Kissinger was a professor of government at Harvard University. As national security adviser and later also secretary of state, Kissinger helped re-establish Sino–US relations in 1972. That meant ending 23 years of animosity over China's turn to Communism and participation in the Korean War, in which US forces had fought to prevent the Communist North Korean state from taking over South Korea.

Kissinger was also an architect of the policy of 'détente' in the cold war with the Soviet Union. Among the measures to 'de-escalate' tensions between the United States and the Soviet Union that he helped put in place was an agreement to limit arms — a dramatic step away from the sort of tensions that had brought the two nations to the brink of nuclear war during the Cuban missile crisis. And by helping to end the Vietnam war, for which he received a Nobel Peace Prize, and paving the way to the Camp David peace accords between Egypt and Israel following war in 1973, Kissinger helped Nixon establish peace in critical conflict zones.

> "Professors should confine themselves to what they know and leave the politics to politicians."

Yet, like two of his immediate academic predecessors as national security adviser, McGeorge Bundy and Walt Rostow, Kissinger also made miscalculations that cost the nation blood, treasure and prestige. At the start of Nixon's term in 1969, Kissinger supported the president's decision to keep troops in Vietnam. Staying in the war brought an additional 23,000 military deaths and failed to save Saigon from a North Vietnamese conquest in 1975. Likewise, Kissinger's collaboration with Nixon in helping the Chilean military topple democratically elected Salvador Allende undermined US standing across Latin America and opened the way to Augusto Pinochet's 17-year dictatorship[3].

Earlier errors of judgement by Bundy and Rostow should have been cautionary tales for Kissinger. Bundy joined John F. Kennedy's administration as national security adviser in 1961. This was a post Rostow later assumed during Lyndon Johnson's presidency after Bundy disagreed with the president on how to encourage public backing of the Vietnam war. A professor of government and the youngest dean of faculty in Harvard's history, Bundy was described by *Washington Post* columnist Joseph Kraft as "unmatched" in his ability "to articulate and execute public purposes" and as perhaps the only member of the postwar generation in government to deserve "the statesman's mantle". Rostow did not lag far behind in reputation. An economist at the Massachusetts Institute of Technology in Cambridge and prolific author of studies that shaped public thinking about economic growth, Rostow, like Bundy, seemed a natural fit for the position of national security adviser.

However, both men badly misread the ability of the United States to control events in Vietnam: they believed that US forces could help the South Vietnamese government defeat Communist Viet Cong insurgents backed by North Vietnam and could assure the rise of a democratic government in Saigon.

Even after the deaths of nearly 60,000 US soldiers and a Vietnam unified under Communist control, Rostow would never concede that sending troops to Vietnam had been a mistake. Instead he argued that the war had given other southeast Asian nations time to develop and avoid Communist takeovers[4–6]. Likewise, Kissinger never acknowledged errors in his and Nixon's dealings with Vietnam and Chile — even though he would be hard pressed to find many defenders among historians who have studied them.

By contrast, Bundy shared former defence secretary Robert McNamara's retrospective conviction about the war that "we were wrong, terribly wrong". Indeed, towards the end of his life Bundy wondered how someone as learned as himself could have been so mistaken[7].

A look ahead

Among the leading academic lights Obama has chosen to join his administration is Steven Chu. As well as Obama's energy secretary, Chu is a Nobel-prizewinning physicist and former head of the Department of Energy's Lawrence Berkeley National Laboratory in California. John Holdren, a Harvard professor of environmental science, is Obama's science adviser. Harold Varmus, a

ILLUSTRATION BY D. THOMPSON

Nobel laureate in medicine, former head of the National Institutes of Health and head of the Memorial Sloan–Kettering Cancer Center in New York, is the chairman of his campaign science advisory council. And Lawrence Summers, the former US Treasury secretary and president of Harvard, heads the White House economic council. What should these advisers, and others besides them, learn from the successes and failures of their predecessors?

The principal lesson I see in assessing the records of intellectually brilliant men such as Oppenheimer, Bundy, Rostow and Kissinger is that academics should always provide advice based on the best available evidence and try not to be swayed by lobbying, or by political or ideological considerations. Total abstinence from politics is not an option, especially for a secretary of energy or a secretary of state who have to take account of both domestic and international political cross-currents, or groups and nations pressing their special interests. Nevertheless, allowing political judgements to overshadow evidence-based understanding is a prescription for making the sorts of errors that are all too common among partisans elected to high offices.

Oppenheimer largely avoided this mistake. He called the bomb "an evil thing" that in time might lead "mankind to curse the names of Los Alamos and Hiroshima". Although he had doubts about building such a destructive weapon, he never allowed his political concerns to interfere with his work.

Bundy and Rostow were different. In heeding political pressures in the White House, they deserted their understanding of how history works. Both advisers believed that a Communist victory in South Vietnam would not only jeopardize Johnson's domestic political standing but also US interests in southeast Asia and Europe, where they feared the Soviets might be emboldened to commit acts of aggression that could threaten a wider war.

The Democratic administration's Bundy and Rostow lived in the shadow of Senator Joseph McCarthy and other right-wing critics, who had pilloried Harry Truman and the Democrats for having 'lost' China to Communism by failing to give sufficient backing to Chiang Kai-shek's nationalist government. Yet their academic expertise should have told them that world events do not simply replicate themselves in vastly different contexts. Vietnam wasn't China. In addition, McCarthyism had lost favour by the 1960s. What's more, there was nothing to suggest that a Communist victory in Vietnam would have any significant effect on the actions of the Soviet Union or China, or on the outcome of the larger cold war.

By focusing so much attention on unrealistic political fears that were largely confined to the White House, Bundy and Rostow encouraged policies that ill-served the United States. They would have served the country better had they devoted more of their efforts to assessing, using the best available knowledge, the likely effectiveness of bombing and ground combat in Vietnam, where the prospects for success were highly questionable.

Kissinger made similar misjudgements. He feared that the collapse of South Vietnam and the continuing control of Chile by a left-wing government would undermine US credibility with both allies and adversaries, and make Nixon vulnerable to charges of having failed to meet the Communist threat in southeast Asia and the western hemisphere. But in 1961, when he became national security adviser, more than three years of US participation in the Vietnam fighting had undermined his country's credibility with both allies and adversaries abroad, not enhanced it. Meanwhile, Fidel Castro's Communist regime in Cuba had turned out to have only limited effect on the United States. This knowledge should have persuaded Kissinger, who prided himself on his standing as a foreign policy realist, to make a quick end to the war and to realize that Allende presented no significant threat to the United States.

Kissinger's deep ties with Nixon almost certainly influenced his thinking. In 1970, Nixon's chief of staff H. R. Haldeman told Kissinger that the president intended to end the US military presence in Vietnam in 1971. Kissinger warned Nixon that if South Vietnam then became unglued, it could jeopardize his re-election in the following year by opening him to attacks for having failed to bring "peace with honour", as he had promised. Instead of making the sort of rigorous calculations about foreign threats a national security official is charged with, Kissinger included domestic political considerations in his advice.

The White House professor may sincerely believe that promoting a president's political standing is vital to the national well-being, but becoming a partisan advocate can be a formula for providing poor advice. In short, professors should confine themselves to what they know and leave the politics to politicians. ■

Robert Dallek is professor of history emeritus at the University of California, Los Angeles. He is the author of *Nixon and Kissinger: Partners in Power* (2007) and *John F. Kennedy: An Unfinished Life* (2003).
e-mail: rdallek@aol.com

1. MacGregor Burns, J. *Roosevelt: The Lion and the Fox* (Harcourt Brace Jovanovich, 1956).
2. Bird, K. & Sherwin, M. J. *American Prometheus: The Triumph and Tragedy of J. Robert Oppenheimer* (Alfred A. Knopf, 2005).
3. Dallek, R. *Nixon and Kissinger: Partners in Power* (Allen Lane, 2007).
4. Dallek, R. *John F. Kennedy: An Unfinished Life, 1917–1963* (Allen Lane, 2003).
5. Dallek, R. *Flawed Giant: Lyndon Johnson and His Times, 1961–1973* (Oxford Univ. Press, 1998).
6. Goldstein, G. M. *Lessons in Disaster: McGeorge Bundy and the Path to War in Vietnam* (Times Books, 2008).
7. McNamara, R. *In Retrospect: The Tragedy and Lessons of Vietnam* (Times Books, 1995).

BOOKS & ARTS

Keeping up with the nuclear neighbours

Since acquiring atomic weapons, India, Pakistan and North Korea have not engaged in major warfare. But nuclear deterrence alone does not buy peace — diplomacy must keep the balance, says **George Perkovich**.

The Long Shadow: Nuclear Weapons and Security in 21st Century Asia
Edited by Muthiah Alagappa
Stanford University Press: 2008. 592 pp.
$75 (hbk), $29.95 (pbk)

South Korea fears a possible nuclear threat from its neighbour, and sees US negotiators as crucial players.

JUNG YEON-JE/AFP/GETTY

The cold war distorted definitions of 'normal' nuclear behaviour. The giant antagonists, the United States and the Soviet Union, built gargantuan arsenals poised for launch at a moment's notice. They poked and prodded each other until the Cuban missile crisis of 1962 chastened them to give arms control a chance. Notwithstanding a series of treaties meant to manage their nuclear competition and help shape a global nuclear order — from the Partial Test Ban Treaty in 1963 through to the Strategic Arms Reduction Treaty II 30 years later — Washington DC and Moscow ordered the construction of thousands more nuclear weapons and kept them ready for use, even when no crisis was at hand.

By the mid-1970s, China, Israel and India had nuclear explosives, and Pakistan and South Africa were preparing to join them. These nations treated nuclear weapons differently. They built relatively few, did not deploy them for immediate use and kept them largely out of political view. South Africa disarmed in the early 1990s, and North Korea became nuclear-armed. Of the nine countries that have nuclear weapons today, the United States and Russia are hardly typical.

The Long Shadow illuminates the different ways that nuclear-armed states have sought to extract the benefits of nuclear weapons while minimizing their risks. Muthiah Alagappa has masterfully edited 14 chapters by leading experts covering the United States, Russia, China, India, Pakistan, Israel, North Korea, Iran, Japan, South Korea, Taiwan and Australia, and, more broadly, the Association of Southeast Asian Nations and the prospects of nuclear terrorism in Asia. Alagappa frames and then interprets these chapters with two of his own. Some chapters are superb, the rest are good. None is bad.

> "The threats of direct conflict are low, but concerns about the nuclear future are high."

The book suffers from a stretched definition of Asia — from Israel eastwards, through the United States, north to Russia and south to Australia — that includes all nuclear-armed states except France and the United Kingdom. Jamming so many countries into one regional construct is unhelpful. Alagappa betrays the problem when he repeatedly generalizes about Asia but then adds that Iran and the Middle East depart from whatever pattern he is describing. The chapters on Iran and Israel, by Devin Hagerty and Avner Cohen respectively, are solid. But they don't add much. Whether the Middle East nuclear challenge ends in disaster or security will depend more heavily on factors other than the nuclear policies of the Asian states to the east.

Asian states have not engaged in major warfare since 1979, which is before India, Pakistan and North Korea acquired nuclear weapons. Alagappa extols the peaceable effects of nuclear deterrence, but it is not clear that deterrence has caused the relative absence of hostilities. With so few threats of direct conflict, the need for nuclear deterrence as a military tool has been low. Indeed, contrary to Alagappa's nuclear bullishness, the nuclear programmes of North Korea, Pakistan and Iran have caused more insecurity than they have alleviated.

Worldwide, there are only three sources of conflict with pressing probabilities of nuclear escalation — between the United States and China over Taiwan, between India and Pakistan and between Iran and the United States or Israel. In each, as Alagappa recognizes, "nuclear deterrence today operates largely in a condition of asymmetric power relationships". Nuclear weapons may partially equalize the military balance of power between states, but this "benefit" is circumscribed. Behaving aggressively behind a putative nuclear shield to change a regional balance would invite other powers "to resort to full-scale conventional retaliation. The onus of escalation to the nuclear level then shifts to the conventionally weaker, revisionist state that initiated the crisis … there is no certainty that international diplomatic intervention would favor the revisionist state."

The India–Pakistan nuclear relationship often

produces intense international hand-wringing. Danger does lurk there, largely owing to Pakistan's political crisis and reluctance to formalize the territorial status quo with India. Stimuli for conflict emerge from Pakistan; competitive logic and political imperatives may lead both states to brinkmanship. As suggested in the chapters on India, by Rajesh Rajagopalan, and on Pakistan, by Feroz Hassan Khan and Peter Lavoy, both countries recognize that nuclear weapons make a war between them unwinnable. Yet they remain unable to transform this recognition into a confident peace that would empower Pakistan's civilian leaders to press the army and intelligence services to concentrate on internal security rather than nurturing low-intensity violence in India and Afghanistan.

The comparative advantage of *The Long Shadow* emanates from the chapters on Japan, China, South Korea and North Korea. Paradoxically, in northeast Asia the threats of direct conflict are low, but concerns about the nuclear future are high. This suggests the political, more than the specifically military, importance of these weapons.

Michael Green and Katsuhisa Furukawa write in the book that nuclear weapons are increasingly present in Japanese thinking, but not as war-fighting instruments or protection against existential threat. "Rather, it is the specter of political and strategic entropy that would be associated with a collapse of the US extended deterrence commitment that is animating strategic thinking in Japan." North Korea's bomb and improved Chinese capabilities reopen "the old question of whether the United States would protect Japan even at the risk of inviting nuclear strikes against US cities". Some Japanese strategic thinkers worry that the United States might "conclude a bilateral arms control agreement with Beijing that endorses protection of Chinese limited nuclear strike capability against the US". They fear this would decouple the United States from Japan.

Kang Choi and Joon-Sung Park describe how South Koreans have an "excessive fear of nuclear threat" combined with a "fear of abandonment" by the United States, and its opposite, "fear of entrapment". They argue that South Korea's fear of abandonment "could soar if the United States tacitly accepted North Korea's nuclear weapon status". Conversely, the fear of entrapment "would linger as long as the public believes that a US military strike on North Korea is possible".

Doubts about the credibility of extended deterrence were much greater during the cold war, as Green and Furukawa and Choi and Park document. Still, policy-makers in Washington, Tokyo, Seoul and Beijing must undertake concerted diplomacy to instil political-strategic confidence in the region in ways that reduce rather than raise the salience of nuclear weapons.

The Long Shadow offers useful guidance to this end. None of the authors urges US retrenchment from the region or rethinking of Japanese, South Korean or Taiwanese nuclear abstinence. Acquisition of nuclear weapons by these countries would only exacerbate insecurity and reduce US commitments to act to defend peace and stability there. Instead, greater effort must be made to enhance the transparency of intentions and capabilities, bolster conventional deterrence and foster unity in dealing with North Korea.

Leaders in the United States and China together hold a key. China will not become more cooperative and transparent and limit its strategic build-up if the United States does not clarify that it is prepared to accept China's nuclear deterrent. This would mean limiting missile defences and certain non-nuclear strike capabilities. Sino-American strategic accommodation need not devalue the US extended deterrent, as some in Japan may fear. As long as nuclear weapons remain, the United States will extend its deterrence umbrella to its allies. To reassure Japan of this, leaders in Washington, Beijing and Tokyo must undertake more forthright strategic dialogues. Framing such dialogue with an explicit objective of creating conditions for incremental, verifiable steps towards nuclear disarmament would add an important Asian dimension to the global effort to live up to the promise made in the 1968 Nuclear Nonproliferation Treaty, the future of which has come into question.

The shadow in this volume's title refers to the chastening threat of nuclear war. The complexity and particularity of the nuclear story in each country surveyed reminds us that the people responsible for preventing the darkness of nuclear war would benefit from the light that careful scholarship can provide. The illumination offered in *The Long Shadow* should be welcomed. ■

George Perkovich is vice-president for studies at the Carnegie Endowment for International Peace and is a co-editor of the book *Abolishing Nuclear Weapons: A Debate*.
e-mail: gperkovich@carnegieendowment.org

Pugwash, nukes and peace

After years of backsliding on nuclear-weapons proliferation by the world's superpowers, President Barack Obama has stated that he intends to "make the goal of eliminating all nuclear weapons a central element" in nuclear policy. His recently appointed chief science adviser, physicist John Holdren, spent ten years as chairman of the executive committee for the Pugwash Conferences on Science and World Affairs, the peripatetic annual meeting of scientists and statesmen to discuss ways to control nuclear weapons. It is named after the Canadian village of Pugwash, Nova Scotia, where its first conference was held under the sponsorship of a wealthy Canadian philanthropist, Cyrus Eaton.

The Strangest Dream
Film directed by Eric Bednarski
Produced by the National Film Board of Canada

Joseph Rotblat: A Man of Conscience in the Nuclear Age
by Martin Underwood
Sussex Academic Press: 2009. 144 pp. £17.95

Professor Pugwash, The Man Who Fought Nukes: The Life of Sir Joseph Rotblat
by Kit Hill
Ryelands: 2008. 80 pp. £8.99

The late Joseph Rotblat would have been heartened by these recent political developments. Rotblat was the youngest signatory of the 1955 Russell–Einstein Manifesto against nuclear weapons, which gave rise to the first Pugwash Conference at the height of the cold war in 1957. Rotblat dedicated more than half a century to the fight to abolish nuclear weapons. In 1995, he and the Pugwash organization shared the Nobel Peace Prize.

Two edited collections on Rotblat were published soon after his death in 2005 at the age of 96. As yet there is no substantial biography, although one is being prepared by the writer Andrew Brown. Now, Rotblat is the focus of *The Strangest Dream* — a Canadian documentary film (http://tinyurl.com/cnehl3) made to celebrate the centenary of his birth — which is intelligent, vivid and all the more powerful for its restraint; and the subject of two brief but interesting books — Martin Underwood's *Joseph Rotblat* and Kit Hill's

Professor Pugwash, The Man Who Fought Nukes. Both authors are physicists who knew Rotblat personally. Hill is a long-standing collaborator in British Pugwash, as mentioned in the foreword by UK Astronomer Royal Martin Rees. Underwood worked as a postdoc with Rotblat on the linear accelerator at St Bartholomew's Hospital in London. Their books aim to introduce Rotblat's life and work to distinct readerships — with uneven results. Ironically, it is the director of the film, Eric Bednarski, who, despite having missed meeting his subject in the flesh, brings Rotblat alive.

Rotblat's first words on screen express his attitude to his science. Speaking in the precise, Polish-accented English he learned in wartime Britain in his thirties, he says: "If my work is going to be applied, I would like myself to decide *how* it will be applied." Not for Rotblat the seductive idea that scientists have no responsibility for the uses to which their discoveries are put. Ethics were as important to him as experiments.

Born in 1908 into a religious Jewish family in Warsaw, reduced to penury by the First World War, Rotblat was forced to become an electrician after leaving school. Eventually he entered academic physics through evening school, worked under a professor trained by Marie Curie and, in mid-1939, left Poland for the University of Liverpool, UK, to conduct nuclear-physics research under James Chadwick, discoverer of the neutron. Atomic fission had just been discovered in Germany, and even before leaving Poland, Rotblat had privately visualized that fission could lead to an atomic bomb. Wrestling with his conscience — like Albert Einstein in 1939 — and leaving behind his Polish wife, who was eventually sent to a Nazi death camp, he decided that he must work on the bomb in case the Germans built one first and won the war. Chadwick, at first reticent to discuss such a sensitive subject with an 'alien', however friendly and able, finally got permission to bring Rotblat to join his team at the Atomic Research Laboratory in Los Alamos, New Mexico — the Manhattan project.

Rotblat was the sole physicist to leave Los Alamos on grounds of conscience before the atomic bomb was dropped on Japan in August 1945. At a dinner party in 1944, he learned from the US army general in charge of the Manhattan project that the real target was Russia, and from Chadwick that Nazi Germany had abandoned its rival project. He resigned immediately and returned to the United Kingdom under a cloud of suspicion from US intelligence that he was a spy for the Soviet Union. A trunk of his papers mysteriously disappeared in transit from Los Alamos, presumably into the archives of the Federal Bureau of Investigation. Some other bomb-making physicists felt qualms in 1945 and even protested to the authorities, but only Rotblat had the "courage" to risk his career for his convictions, observes Pakistan Pugwash nuclear physicist Pervez Hoodbhoy in the film. "He was not the kind of man to be told what to think," says Rotblat's Polish niece Halina Sand.

Joseph Rotblat won a Nobel prize for his work on nuclear disarmament with the Pugwash organization.

J. EGGITT/AFP/GETTY

> "Only Joseph Rotblat had the courage to risk his career for his convictions."

This is mainly why Pugwash was effective during the cold war. The first conference was attended by one lawyer and 21 scientists from the United States, the Soviet Union, the United Kingdom, China, France, Poland, Australia, Japan, Austria and Canada. Despite pressure from governments, Rotblat and the Pugwash Conferences refused to toe official lines. Instead, participants — whether Soviet scientists or statesmen such as former US defence secretary Robert McNamara — spoke as individuals. The meetings were private, but not secret, and held without the presence of the media. Formal speeches were generally eschewed; discussions took place around a table and informally, with the agreement that contributions would not be publicly attributed to individuals, so they could speak relatively freely. The result, notes Underwood, is that Pugwash was instrumental in achieving the signing of the Partial Test Ban Treaty in 1963 and, in 1972, both the Biological Weapons Convention and the Anti-Ballistic Missile Treaty. It also helped mediate between Moscow and Washington DC during the Cuban missile crisis of 1962, and established strong links with the Soviet leader Mikhail Gorbachev, who admired Rotblat, at the time of Gorbachev's arms negotiations with US President Ronald Reagan in the 1980s.

Underwood emphasizes politics more than science, and writes conventionally. Hill is more impressionistic and quirky, with the science explained at a very basic level in boxes. Both books contain errors; for example, Marie Curie's second Nobel prize was not for work on "artificial radioactivity" done with her daughter, as claimed by Hill. But it is nice to know from his book that Captain Pugwash, the British comic-strip pirate created in 1950 — whose fame initially made Rotblat suspect that Eaton's offer of sponsorship was a hoax — later sent the Pugwash Conferences a congratulatory scroll. ■

Andrew Robinson is a visiting fellow of Wolfson College, University of Cambridge, Cambridge CB3 9BB, UK. His book *Einstein: A Hundred Years of Relativity* contains material by Joseph Rotblat on Einstein's quest for global peace.
e-mail: ar471@cam.ac.uk

Q&A: The art of transmutation

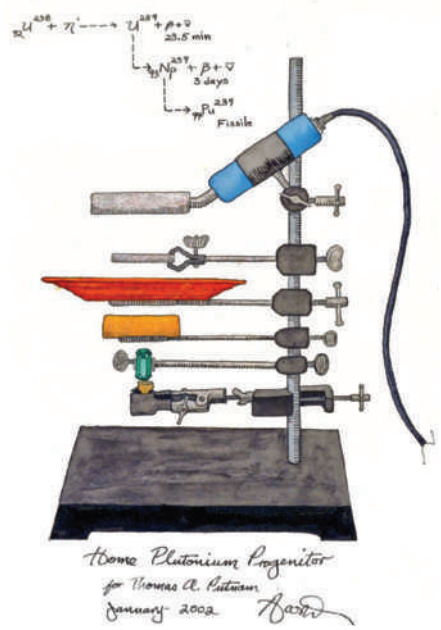
James Acord is the only sculptor licensed to work with radioactive materials. Formally trained in nuclear physics, he tells *Nature* why he thinks contaminated nuclear sites should be marked for future generations and explains his obsession with the nuclear age.

Why do you think nuclear sites should be marked?
The land around the decommissioned US nuclear-processing facility in Hanford, Washington state, is so contaminated that it will never be completely cleaned up. Far into the future the site should carry warnings that transcend changes in language and society to discourage people from growing crops there. I would love to produce something of lasting aesthetic significance, both as a warning marker and as a commemorative piece for the advent of the nuclear age. I lived at the site for 7 years but I never really fitted in. Being an artist made me an outsider in a largely engineering and scientific community. I live in Seattle now.

A. S. AUBRY

Sculptor James Acord wants to create artwork to mark the US facility that first produced plutonium for weapons.

What are you working on?
My attention is on a device here in my studio that transmutes uranium to plutonium. It symbolizes the process that produced the plutonium for the first nuclear-weapon test during the Second World War. Transmutation, which I define as changing the number of protons in any atomic element, is an inevitable tool of sculpture — altering one material into another. In the device (sketched below), I've taken the radioactive element americium — a source of α-radiation — out of dismantled smoke detectors and put it in contact with a small emerald, which converts the α-particles into neutrons. A six-centimetre-thick slab of beeswax then serves as a hydrogen moderator, increasing the chances of transmutation. The neutrons coming out of the beeswax filter go into triuranium octoxide, which is found in a red glaze used in 1940s ceramics. Some of the uranium in the glaze will become plutonium.



J. ACORD

Has your work been well received by the nuclear industry?
I confess that I'm disappointed and surprised at how little support I've had. Some of my ideas, such as finding a way of transmuting the element technetium-43 to ruthenium-44, are as great as sliced bread and I don't see why the people running nuclear reactors won't invite me in to use them. There's a sense in the scientific and engineering community that artistic use of the nuclear process is frivolous. But this makes me more determined.

Why were you blocked from using a reactor while an artist-in-residence?
During my 1998–99 residency at Imperial College London, my goal was to use their reactor to transmute technetium to ruthenium. However, while I was there, a couple of nuclear accidents occurred in the world. A senior member of staff said I couldn't use it: 'absolutely not, we don't want any publicity; we don't want Londoners to know we're operating a nuclear reactor in the city'. I got my nose out of joint about it and made some metal sculptures that said in gold foil 'no access'.

What do you think about nuclear politics?
There's no reason for me to be either pro- or anti-nuclear. We are in a nuclear age, for good or for ill. The physics of the nuclear age is unmistakable and we'll have to embrace more nuclear energy in the future. But the number of people making decisions about it is extremely small. Sculpture, which is an art of technology, should be free to address the technology that is characteristic of our time.

Is your sculpture safe even though it's made from radioactive materials?
Do I think it's safe? Yes I do. Is it legally safe? I'm not so sure. The piece I'm doing now, strictly speaking, is not covered by my licence. The finished work of art, containing both uranium and plutonium, will be slightly radioactive. When I first began removing the uranium-bearing glaze from ceramic tableware, I wasn't very careful about dust inhalation. I suppose it has increased my chance of lung cancer. But at my age I don't worry. Sculpture is a hazardous profession. ■

Interview by **Daniel Cressey**, a reporter for *Nature*.

NEWS & VIEWS

ECOLOGY

Gini in the bottle

Shahid Naeem

An elaborate microcosm study has a message for the wider world: declining distributional equity among species, where the rare become rarer, and the dominant become more dominant, can put ecosystems at risk.

In the 1770s, Joseph Priestley, the father of biogeochemistry[1], conducted his famous experiments in which he placed mice and mint plants in bottles, and discovered the balance between 'putrefying' and 'regenerative' processes. Priestley thus began the tradition of using organisms in microcosms to explore nature. He and his colleagues clearly recognized the global significance of his findings, despite their small scale. On page 623 of this issue, Wittebolle *et al.*[2] describe ecological research in that tradition, but carried out with twenty-first-century tools.

Life on Earth is more than mice and mint, of course, with most ecosystems containing hundreds to thousands of plant, animal and microbial species. The balance between 'putrefying' and 'regenerative' processes in nature is better known today as the balance between respiration, decomposition, photosynthesis, primary production and many other ecosystem functions carried out by species that cycle matter between inorganic and organic forms. In the 1990s, motivated by growing concern over dramatic declines in biological diversity, ecologists began to test experimentally whether declining biodiversity — species richness — could adversely affect ecosystem functions. Early studies consisted of an eclectic mix of experiments, manipulating, for example, the richness of microbial species in Petri dishes or bottles, the richness of plant and animal species in growth chambers or artificial ponds, or the richness of plant species in grassland plots.

These investigations were surrounded by controversy, but most of them indicated that ecosystem functions, such as respiration and primary production, were indeed adversely affected by dramatic declines in biodiversity. Today, evidence that plant, animal and microbial biodiversity influences terrestrial[3] and marine ecosystem functions[4] has been documented by hundreds of studies. Such research, however, has focused on changes in species richness, in spite of the fact that changes in the relative abundance of species, or species evenness, are more prevalent and more likely to affect ecosystem function[5,6]. Wittebolle *et al.*[2] now report on what is probably the most elaborate microcosm study ever conducted to examine the influence of biodiversity on ecosystem function. It is devoted entirely to evenness.

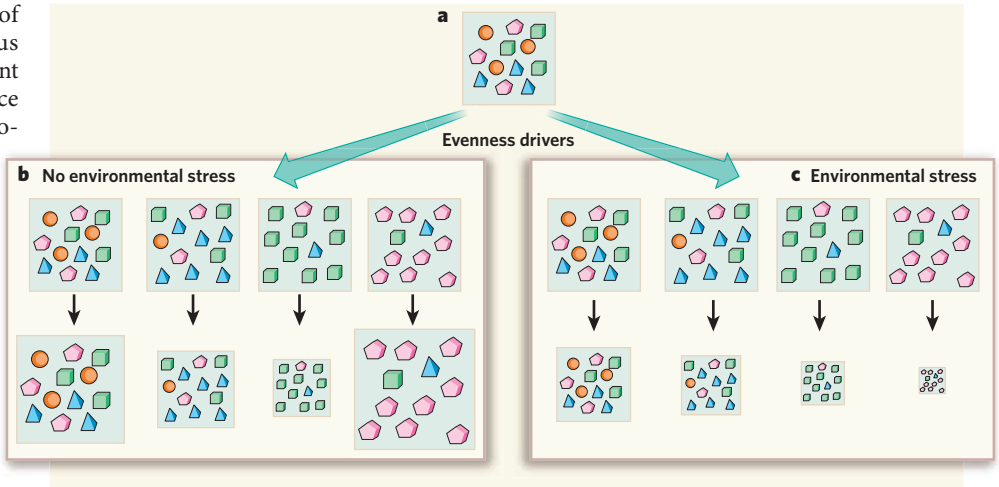


Figure 1 | Implications of Wittebolle and colleagues' results[2]. **a**, Relative species abundances in an initial community with maximum evenness, shown by the symbols, may alter in response to 'evenness drivers'. Conservation, for example, may enhance evenness, whereas selective harvesting, species displacement by invasion, and agriculture reduce it. In consequence, ecosystem function will decrease (smaller square) or increase (larger square). **b**, With no environmental stress, highly uneven communities, such as agricultural systems, may exhibit high levels of ecosystem function, as shown by the largest square. Otherwise, a decline in functions accompanies declining evenness. **c**, Environmental stress, such as a change in temperature, pH or salinity, reduces ecosystem functioning with increasing severity as evenness declines. The smallest square depicts a system in which the dominant species was the most sensitive to the stress, and stress-resistant species were rare.

Manipulating richness is logistically challenging but straightforward; manipulating evenness is not. To manipulate richness, one constructs replicate ecosystems that vary in numbers of species while holding the total number of individuals constant and distributing individuals equally among species. For example, in an ecosystem that could be made up of three species totalling 300 individuals, one would manipulate richness by constructing replicates that contain 300 individuals of a single species, or 150 individuals each of two species, or 100 individuals each of three species. In contrast, to manipulate evenness, distributional equity is varied. Thus, with three species, one would construct replicates containing 100, 100 and 100 individuals (known as perfect equity), 99, 99 and 102 individuals, 98, 99 and 103 individuals, and so on until reaching 1, 1 and 298 (near perfect inequity). As such a complete experiment is not practical, a set of replicates is instead constructed that represents a comprehensive, unbiased sampling of possible abundance distributions. How to construct such a set is a considerable challenge in the study of evenness. Wittebolle *et al.*[2] found an elegant solution based on a widely used metric of distributional equity; nevertheless, implementing it still required a staggering 1,260 microcosms.

There are many metrics of evenness, each of which has its pros and cons[7–9]. Of these metrics, Wittebolle *et al.* chose the Gini coefficient (G), whose virtue is that it is based on the Lorenz curve, a graphical representation that neatly describes distribution equity as the relationship between the cumulative proportion of species richness and the cumulative proportion of species abundance. Every community's relative abundance can be described by a Lorenz curve; G is simply the area of the region

bounded by this curve and the straight-line diagonal describing perfect equity (see Fig. 3 of Wittebolle and colleagues' Supplementary Information[2]).

Rather than mice and mint, Wittebolle *et al.* used bacterial species, which meant that their microcosms could be small — indeed, as wells in microplates, they were *very* small. Each well contained 18 denitrifying species (largely proteobacteria), at densities of 10^7 per millilitre. Denitrifying bacteria metabolize nitrates and nitrites, and the level of denitrification provided a measure of ecosystem function. With the aid of modern tools to assess net denitrification, such as flow cytometry, ultracold freezers, robot pipetters and spectrophotometric microplate readers, the microplate system made exploring evenness possible at a level of thoroughness simply unimaginable by more typical ecological methods.

The thoroughness of this study[2] makes its results rather convincing. The authors found that declining evenness affects ecosystem functioning in much the same way as declining richness does. But the magnitude of the impact depends on the nature of the stress the ecosystem is experiencing and the functional traits of the dominant species (such traits are the properties that govern how species respond to or affect their environment, in this case[2] tolerance to cold or salinity; Fig. 1). For example, no bacterial species fared well when microcosms were exposed to cold stress. When exposed to salinity stress, however, some species were more salt tolerant than others; thus, microcosms with greater evenness were more likely to have enough salt-tolerant individuals to assure net denitrification.

These findings do not mean that we should run out and increase species evenness. Natural ecosystems are typically uneven, but the real world is highly heterogeneous, spatially and temporally, unlike the highly controlled conditions of this study. In the real world, different species will naturally dominate in different places and at different times, so the potential value of rare species is missed in studies where conditions do not fluctuate. There is also a growing literature suggesting that the richness and evenness of functional traits[10] are more relevant to ecosystem functioning than species richness and evenness. What mattered in this study, for example, was the diversity of stress-tolerance traits, not the species diversity. The real world is also trophically complex, making one wonder what the results might have been if viruses or microflagellates that prey on the bacteria had been present. These are directions future research should take; but as the level of detail in the authors' supplementary material illustrates, to do that will be daunting.

Do we need to go much further, however, before delivering the clear message of this research? Wittebolle and colleagues' study is technically sophisticated, abstract and small in scale. Nonetheless, the implications are global, much as Priestley's message about the balance of nature was more than two centuries ago. Ecosystems worldwide are becoming dominated by one or a few domesticated or invasive species[11]. So it seems likely that ecosystem functions and the services they provide are becoming less and less resilient to the stresses, such as climate change, nitrogen deposition and salt-water intrusion, that are being generated by the world's rapidly increasing population. ■

Shahid Naeem is in the Department of Ecology, Evolution, and Environmental Biology, Columbia University in the City of New York, New York, New York 10027, USA.
e-mail: sn2121@columbia.edu

1. Gorham, E. *Biogeochemistry* **13,** 199–239 (1991).
2. Wittebolle, L. *et al. Nature* **458,** 623–626 (2009).
3. Cardinale, B. J. *et al. Nature* **443,** 989–992 (2006).
4. Worm, B. *et al. Science* **314,** 787–790 (2006).
5. Wilsey, B. J. & Potvin, C. *Ecology* **81,** 887–892 (2000).
6. Hillebrand, H., Bennett, D. M. & Cadotte, M. W. *Ecology* **89,** 1510–1520 (2008).
7. Smith, M. D. *et al. Oikos* **106,** 253–262 (2004).
8. Buzas, M. A. & Hayek, L.-A. C. *Paleobiology* **31,** 199–220 (2005).
9. Gosselin, F. *J. Theor. Biol.* **242,** 591–597 (2006).
10. McGill, B. J. *et al. Trends Ecol. Evol.* **21,** 178–185 (2006).
11. Millennium Ecosystem Assessment *Biodiversity Synthesis Report* (Island, 2005).

SOLID-STATE PHYSICS

Spin's lifetime extended

Jaroslav Fabian

Electrons in semiconductors are subject to forces that make their spins flip. According to new evidence, if an ensemble of spins curls into a helix, the collective spin lifetime can be greatly enhanced.

Over the past decade, electron spin — the electron's intrinsic rotation, which is commonly described as 'up' and 'down' and which gives rise to its magnetic moment — has come to the forefront of research in solid-state physics. A whole new field, called spintronics[1–4], has emerged as an umbrella for both applied and fundamental research on spin transport and spin control in metals and semiconductors. On the applied front, spintronics is already realizing its potential in applications such as magnetic read heads in computers' hard disks or magnetic random-access memories that are non-volatile — that is, they can retain information even when the power is turned off. On the fundamental side, the field is generating equally fascinating discoveries of spin phenomena. One such discovery, the realization of a 'persistent spin helix' in a semiconductor is reported by Koralek and colleagues[5] on page 610 of this issue.

Spin is an intrinsic property of the electron that never goes away. But unlike the electron charge it has two possible values, positive (up) and negative (down), which are linked to the spin-axis orientation. This means that the net spin of an ensemble of electrons can decay. Start with an ensemble of spin-up electrons and in a nanosecond or so you may find that they are equally 'up' and 'down', resulting in an ensemble that has no net spin.

In semiconductors, the major cause of spin decay is a rather weak, and up to recently underappreciated, quantum interaction called spin–orbit coupling. This interaction couples the electron velocity (orbit) with the electron spin. The electron velocity changes randomly when the electron moves past imperfections in the semiconductor's crystal structure or changes simply as a result of atomic-lattice vibrations. Because of spin–orbit coupling, the spin orientation of the electron changes as well. But only a little: the electron needs thousands or even millions of velocity kicks, depending on the semiconductor, for its spin to flip and erase the memory of its original orientation.

In spintronics applications, long — tens to hundreds of nanoseconds — spin relaxation times (the time it takes an itinerant electron to flip its spin) are desired to preserve the information encoded in the spin as electrons travel through spintronic devices[2]. To inhibit spin relaxation as much as possible, we could envisage eliminating crystal imperfections and atomic vibrations, but this would be a quixotic exercise in fighting the laws of thermodynamics.

In their experiment, Koralek *et al.*[5] focus on spin–orbit coupling instead. Although such coupling cannot be switched off, it can be tailored by tuning the underlying spatial anisotropy of the semiconductor quantum well — a thin layer of semiconductor material (in this case, gallium arsenide sandwiched between two layers of another semiconductor), which restricts the movement of electrons in the dimension perpendicular to the plane of the layer. The quantum well's spatial anisotropy discriminates between two possible spin orientations in the plane of the quantum well. Because of spin–orbit coupling, this anisotropy is reflected in an anisotropy of spin relaxation[6], which has been explored in a spectrum of themes, from spintronic devices[7] to the propagation of plasmons (quanta of electronic plasma oscillations)[8].

By tuning such spatial anisotropy and by curling electron spins into a helical wave of a certain wavelength and pitch, Koralek and colleagues demonstrate that spin relaxation

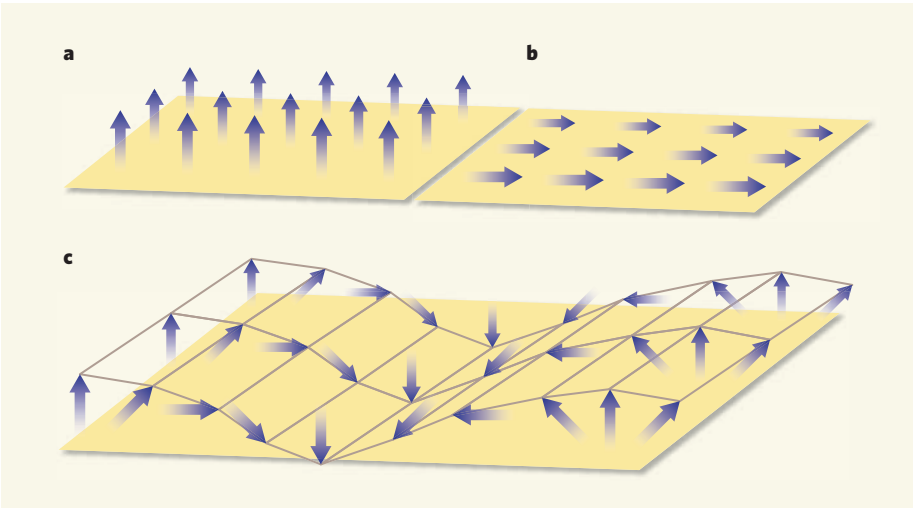


Figure 1 | Persistent spin helix. In a semiconductor quantum well, a thin layer of semiconductor material sandwiched between two other semiconductors, electrons are confined in the dimension perpendicular to the plane of the layer — that is, they move only along the layer (yellow). By a process known as optical orientation[2], electron spins (arrows) can be made to orient out of the plane (**a**) or along one of the plane's dimensions (**b**). In both cases, the electrons are subject to random forces that, in conjunction with an interaction called spin–orbit coupling, cause their spins to flip. Koralek and colleagues[5] show that by combining these two spin orientations to form a helical wave of rotating spin orientation (**c**), and by fine-tuning the structural properties of the quantum well, the spins become largely protected against decay.

becomes inhibited. The authors show that the collective spin-orientation wave persists for much longer than its individual spin components. That is, spins in the helical wave become immune against relaxation: spin–orbit coupling is effectively absent, making the spin unaware of the random velocity kicks. This so-called persistent spin helix, which was theoretically introduced by Bernevig *et al.*[9], is based on a dynamical symmetry of the entire spin ensemble that is formally akin to the rotational symmetry a single electron spin enjoys in the absence of spin–orbit coupling (Fig. 1).

Despite the appeal of the theoretical ideas behind the persistent spin helix, a theorist's notion of fine-tuning spin–orbit coupling and creating rather special spin helices is a far cry from the experimental effort required to realize them. And yet Koralek and colleagues have succeeded in doing just that. Their experimental demonstration of the persistent spin helix is a remarkable feat.

To create spin-orientation waves of the required wavelength, the authors used a technique known as transient spin grating[10,11]. In their experiment, two non-collinear laser beams of light linearly polarized in orthogonal directions interfere at the plane of the quantum well and produce a sinusoidal pattern of light helicity: stripes of alternating circular polarization (helicity) of light. Such a pattern of helicity can orient electrons' spins through a process called optical orientation[2] and generate a spin-orientation wave. Say that right- or left-circularly polarized light creates spin-down and spin-up electrons, respectively. The pattern of light helicity then translates into an identical pattern of electron spin orientation. The wavelength of such a spin-orientation wave can be tuned by changing the angle between the two interfering laser beams.

The resulting (linearly polarized) spin-orientation wave can be viewed as composed of two spin helices — waves of rotating spin orientation. One helix rotates clockwise, the other anticlockwise. Under the right conditions, only one of them is the persistent spin helix. The other helix decays as usual. By watching the temporal evolution of the spin-orientation wave pattern with probe laser beams, we should in principle spot an initial fast decay of the normal (non-persistent) helix, followed by a much slower decay of the persistent one.

This is exactly what Koralek and colleagues find in their experiments. By suitably tuning the structural composition of the quantum wells, achieved by varying both the width and the degree of doping asymmetry of the quantum well, the authors show that the emergent persistent spin helix lasts a hundred times longer than the normal one. The spins curl themselves up to ward off spin relaxation. The slow decay of the persistent spin helix is caused by residual spin–orbit interactions.

Koralek and colleagues' experimental realization of the persistent spin helix is a breakthrough towards minimizing and controlling spin relaxation in electronic systems. The next chapter in the field of spintronics is one that deals with ways of controlling the spin's lifetime electrically. That could be achieved by turning spin helices on and off with an electrical gate, or by demonstrating their role in the predicted drastic increase of electrical spin injection efficiency, an essential part in the operation of spintronic devices[12]. For the spin, this is as good as it gets — at least for now. ■

Jaroslav Fabian is at the Institute for Theoretical Physics, University of Regensburg, 93040 Regensburg, Germany.
e-mail: jaroslav.fabian@physik.uni-regensburg.de

1. Das Sarma, D. *Am. Sci.* **89**, 516–523 (2001).
2. Žutić, I., Fabian, J. & Das Sarma, S. *Rev. Mod. Phys.* **76**, 323–410 (2004).
3. Fabian, J., Matos-Abiague, A., Ertler, C., Stano P. & Žutić, I. *Acta Phys. Slov.* **57**, 565–907 (2007).
4. Awschalom, D. D. & Flatté, M. E. *Nature Phys.* **3**, 153–159 (2007).
5. Koralek, J. D. *et al. Nature* **458**, 610–613 (2009).
6. Averkiev, N. S. & Golub, L. E. *Phys. Rev. B* **60**, 15582 (1999).
7. Schliemann, J., Egues, J. C. & Loss, D. *Phys. Rev. Lett.* **90**, 146801 (2003).
8. Badalyan, S. M., Matos-Abiague, A., Vignale, G. & Fabian, J. Preprint at http://arxiv.org/abs/0804.3366 (2008).
9. Bernevig, B. A., Orenstein, J. O. & Zhang, S.-C. *Phys. Rev. Lett.* **97**, 236601 (2006).
10. Cameron, A. R., Riblet, P. & Miller, A. *Phys. Rev. Lett.* **76**, 4793–4796 (1996).
11. Weber, C. P. *et al. Nature* **437**, 1330–1333 (2005).
12. Cheng, J. L., Wu, M. W. & da Cunha Lima, I. C. *Phys. Rev. B* **75**, 205328 (2007).

DNA REPAIR

New tales of an old tail

Jiri Lukas and Jiri Bartek

Modifications of DNA-associated histone proteins maintain genome integrity. On damage to DNA, phosphorylation of histone H2A.X determines whether repair is justified or if the damaged cell must die.

Chromosomal DNA wraps around histone proteins to form a complex scaffold called chromatin[1]. The reorganization of these proteins following DNA damage is crucial for repairing the damage, and so maintaining genomic integrity and reducing the likelihood of cell death or cancer. One such histone modification — known as γ-H2A.X — follows DNA double-strand breaks (DSBs) and involves phosphorylation by the enzyme ATM of serine residue 139, which is located in the carboxy-terminal tail of the histone variant H2A.X (ref. 2). γ-H2A.X generates a chromosomal microenvironment that promotes recruitment of repair proteins[3] and facilitates DNA repair to reduce the risk of mutations[4]. But how this modification is regulated and how it affects cell fate have remained elusive. Two papers[5,6], including one on page 591 of this issue, provide insights into these questions.

NATURE|Vol 458|2 April 2009

The discovery of DSB-induced γ-H2A.X sparked enormous efforts to decipher how repair and signalling proteins assemble into foci on chromatin marked by this modification. The search concentrated mainly on identifying repair factors and other histone modifications operating downstream of γ-H2A.X — hence 'moving away' from this priming DSB-associated histone mark. But it emerges that another key chromatin modification in response to DSBs also occurs in the H2A.X tail, just three amino acids away from serine 139 (S139).

Indeed, Xiao *et al.*[5] and Cook *et al.*[6] have now independently discovered that tyrosine residue 142 (Y142) of H2A.X is also phosphorylated (Fig. 1a). Both groups show, however, that unlike S139, Y142 is already phosphorylated in unstressed cells and becomes gradually dephosphorylated after DNA damage. Even more unexpectedly, dephosphorylation of Y142 seems to be a prerequisite for the γ-H2A.X modification, indicating that the phosphorylation status of the Y142 residue of H2A.X regulates what has been considered the main trigger of the entire DSB-induced chromatin pathway. Such a twist in our thinking about genome-maintenance mechanisms clearly deserves a closer look.

The starting point for Xiao *et al.*[5] was the observation that the evolutionarily conserved Y142 in human H2A.X is phosphorylated *in vivo*. They then found that components of the WICH chromatin-remodelling complex[7,8] interact with the carboxy terminus of H2A.X, where Y142 is located. Strikingly, they showed that the WSTF component of WICH has tyrosine-kinase activity, enabling it to phosphorylate Y142. The authors also found that, after DNA damage, WSTF dissociates from chromatin — consistent with a decrease in Y142 phosphorylation — making way for the γ-H2A.X modification (Fig. 1b).

Cook *et al.*[6] observed that, during embryonic development of mouse kidney, deletion of either *Eya1* or *Eya3* — genes encoding protein-phosphatase enzymes that dephosphorylate tyrosine residues — coincides with increased γ-H2A.X. The authors also found that the Eya1 and Eya3 enzymes bind to and co-localize with γ-H2A.X at foci of DSBs in the nucleus, leading them to consider that H2A.X might be phosphorylated on a tyrosine residue; indeed, they identified Y142 as the target. What's more, in agreement with the observations of Xiao and colleagues, Cook *et al.* report that after DNA damage there is an Eya1- or Eya3-dependent decrease in tyrosine phosphorylation of H2A.X (Fig. 1b). Reducing Eya levels prevented DNA-damage-induced dephosphorylation of Y142 and the proper interaction of γ-H2A.X with MDC1 — an adaptor protein that senses γ-H2A.X and orchestrates the assembly of repair proteins on the chromatin at DSBs[9,10].

Together, these findings[5,6] make a compelling case for Y142 phosphorylation as a new modification of H2A.X and suggest that a balance between the kinase activity of WSTF and the phosphatase activity of Eya proteins regulates both the formation of γ-H2A.X-marked chromatin and the recruitment of repair factors to DSBs. And, besides uncovering another dimension of the chromatin response to genotoxic stress, each paper provides other surprising results.

First, the WAC catalytic domain that Xiao and colleagues[5] identified in the amino terminus of WSTF shares no sequence similarity with other known kinase enzymes[4] — an intriguing finding, the significance of which extends beyond DNA repair. WSTF probably also phosphorylates substrates other than H2A.X, and the identification of these might help explain the clinical symptoms associated with Williams–Beuren syndrome, a neurodevelopmental disorder linked to deletions of the *WSTF* gene. Furthermore, other proteins might contain a WAC domain, and a search for such hitherto unrecognized tyrosine kinases could be rewarding.

Second, Cook *et al.*[6] report that peptides derived from the carboxy-terminal tail of H2A.X that were phosphorylated on both S139 and Y142 did not bind MDC1, consistent with the fact that Y142 dephosphorylation is required for γ-H2A.X–MDC1 interaction. What was unexpected, however, was that the doubly phosphorylated H2A.X peptide binds the protein kinase JNK1 — an established inducer of programmed cell death (apoptosis). It seems, therefore, that phosphorylated Y142 might function as a decision-maker, determining cell fate after DNA damage. When repair is possible, Y142 is dephosphorylated, allowing the γ-H2A.X modification and the recruitment of repair factors (Fig. 1c). Otherwise, Y142-phosphorylated H2A.X persists, recruiting the JNK1 complex to 'switch' to the pro-apoptotic mode, and eliminate cells with irreversibly damaged genomes from the organism (Fig. 1d).

As with all inspiring discoveries, the work of Xiao *et al.*[5] and Cook *et al.*[6] raises yet more questions. As WSTF is the kinase responsible for Y142 phosphorylation — and could thus be viewed as a negative regulator of γ-H2A.X — one would predict that reducing WSTF levels could facilitate γ-H2A.X formation. In fact, the opposite happens: in the absence of WSTF, γ-H2A.X and focus formation cannot be sustained, and MDC1 recruitment to DSBs is inhibited[5]. To explain this conundrum, Xiao *et al.*[5] propose that WSTF might also help adjust local chromatin structure for maintenance

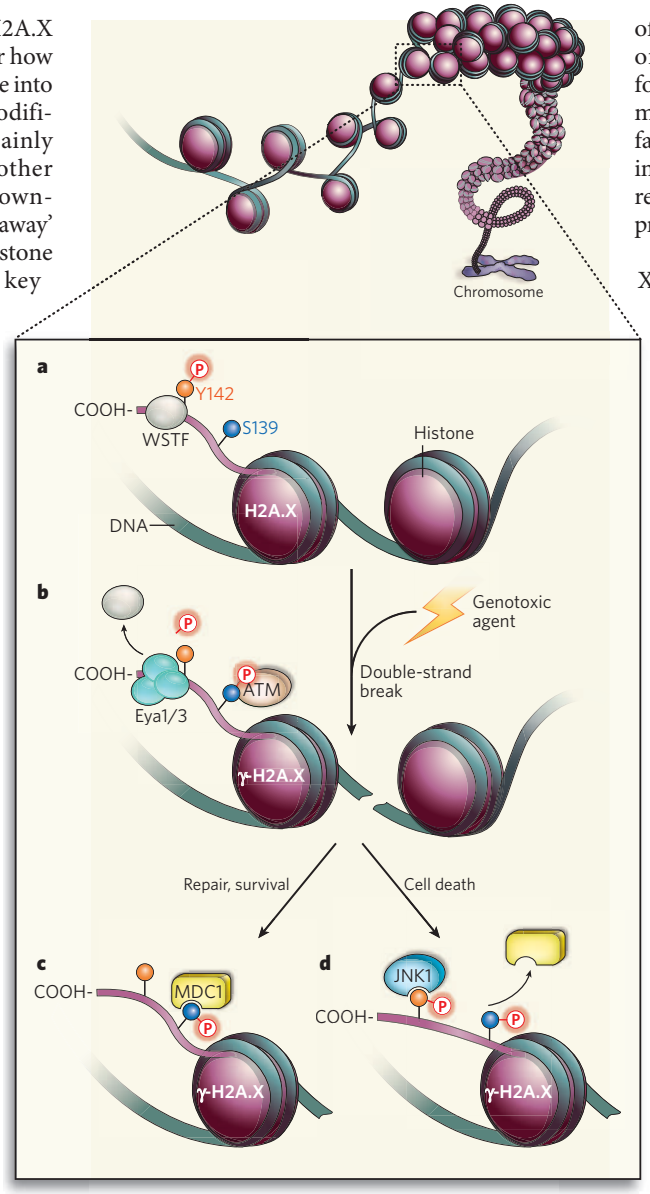


Figure 1 | A matter of life or death[5,6]. **a**, Normally, the WSTF kinase associates with the carboxy terminus of the histone variant H2A.X and phosphorylates its Y142 residue. Thus, chromatin remains in a 'standby' mode with no unnecessary DNA repair events. **b**, When DNA double-strand breaks occur after exposure to genotoxic agents, WSTF dissociates and is replaced with the Eya1/3 phosphatases, which dephosphorylate Y142, facilitating S139 phosphorylation (the γ-H2A.X modification) by the ATM enzyme. What happens next depends on whether the damage is repairable. **c**, If repair is possible, phosphorylated S139 recruits MDC1 and other repair factors. **d**, If it is not, the γ-H2A.X tail might undergo conformational changes that allow maintenance or re-phosphorylation of Y142. This would prevent retention of repair factors, and instead attract the JNK1 complex, which promotes apoptosis.

of γ-H2A.X. This is plausible, as the WICH complex also has chromatin-remodelling activity during DNA replication[7,8].

The main conceptual issue arising from Cook and colleagues' results[6] is the proposed role of phosphorylated Y142 in promoting cell death. On one hand, the authors provide evidence for increased H2A.X–JNK1 interaction in cells exposed to high doses of radiation. This indeed supports the switch model, as such Y142-mediated recruitment of JNK to sites of DSBs helps direct cells towards apoptosis as a last resort. On the other hand, they show that Y142 is dephosphorylated after DNA damage, resulting in the loss of the 'docking site' for JNK1. At first glance at least, this finding does not fit the switch model, calling for more work to reconcile it with the observed pro-apoptotic effects of Y142 phosphorylation. It may be, however, that Y142 is re-phosphorylated after futile attempts to repair excessive DNA damage.

Clearly, the issue of the efficiency of DSB repair and the role of posttranslational chromatin modifications in this process is here to stay. Nevertheless, the two papers[5,6] provide a fresh conceptual framework and tools to tackle this challenge, which should enable us to better understand the genesis of major genome-instability diseases, including cancer, premature ageing and neurodegeneration. ■

Jiri Lukas and Jiri Bartek are at the Institute of Cancer Biology and the Centre for Genotoxic Stress Research, Danish Cancer Society, Strandboulevarden 49, DK-2100 Copenhagen, Denmark.
e-mails: jil@cancer.dk; jb@cancer.dk

1. Groth, A., Rocha, W., Verreault, A. & Almouzni, G. *Cell* **128,** 721–733 (2007).
2. Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. *J. Biol. Chem.* **273,** 5858–5868 (1998).
3. Fernandez-Capetillo, O., Lee, A., Nussenzweig, M. & Nussenzweig, A. *DNA Repair* **3,** 959–967 (2004).
4. Bartek, J. & Lukas, J. *Curr. Opin. Cell Biol.* **19,** 238–245 (2007).
5. Xiao, A. *et al. Nature* **457,** 57–62 (2009).
6. Cook, P. J. *et al. Nature* **458,** 591–596 (2009).
7. Poot, R. A. *et al. Nature Cell Biol.* **6,** 1236–1244 (2004).
8. Bozhenok, L., Wade, P. A. & Varga-Weisz, P. *EMBO J.* **21,** 2231–2241 (2002).
9. Stucki, M. *et al. Cell* **123,** 1213–1226 (2005).
10. Lukas, C. *et al. EMBO J.* **23,** 2674–2683 (2004).

ENVIRONMENTAL SCIENCE

Clean coal and sparkling water

Werner Aeschbach-Hertig

Subsurface storage of carbon dioxide is a major option for mitigating climate change. On one account, much of the gas sequestered in this way would end up as carbonic acid in the pore waters of the host rock.

Atmospheric concentrations of greenhouse gases, especially carbon dioxide, continue to rise at an alarming rate. We seem unable to tame our appetite for fossil fuels on a meaningful timescale, and the concept of carbon capture and storage has emerged as a serious option for reducing CO_2 emissions to the atmosphere. A 'clean coal' technology, in which CO_2 is collected from coal-fired power plants and stored safely below ground, might enable us to continue using this comparatively cheap and abundant energy source without climatic worries.

However, little is known about the long-term fate of large quantities of CO_2 put into geological storage. Gilfillan *et al.*[1] (page 614 of this issue) illuminate this crucial matter by showing that dissolution in groundwater is by far the most important trapping mechanism for CO_2 in the subsurface environment. In other words, sequestering CO_2 in geological formations would probably produce vast quantities of highly CO_2-enriched sparkling water.

The safety of geological storage of CO_2 is obviously a central concern in planning carbon sequestration on a large scale. When CO_2 is injected into the subsurface, it will be retained by physical and geochemical mechanisms[2]. Physical trapping is provided by the presence of sealing, low-permeability rock formations above the targeted layer. Such cap rocks are essential features of natural gas and oil reservoirs, and are a primary requirement for CO_2 storage sites. A further level of safety is added by geochemical interactions that remove the pure CO_2 phase, either through dissolution in water (solubility trapping) or by precipitation of carbonate minerals (mineral trapping). Clearly, mineral trapping is the preferable pathway, as it promises to store the carbon over geological timescales.

To assess the risk of leakage from storage reservoirs, an expansive programme for monitoring underground CO_2 injection in a variety of geological settings has been called for[3]. There are only a few currently active pilot sites, and more are needed. But that apart, such monitoring programmes can reveal the effects of carbon sequestration only on the engineering timescale — they do not yield a direct answer to questions regarding the long-term behaviour of CO_2 in geological storage.

In this respect, the approach taken by Gilfillan *et al.*[1] is logical and informative. The authors used CO_2-rich gas fields as natural analogues for future carbon-storage sites. Other researchers have exploited this idea[4]. But in offering a self-consistent evaluation of noble

50 YEARS AGO

It often happens that investigators, particularly in the social sciences, must try to collect the information which they need by using questionnaires. One of the many problems that are apt to arise concerns the reliability of answers to questions which require an exercise of detailed and specific memory. Recently, the Tobacco Manufacturers Standing Committee issued a Research Paper (No. 2) entitled "The Reliability of Statements about Smoking Habits" by G. F. Todd and J. T. Laws ... The authors show how statements about current smoking habits are generally reconstructed from a sort of 'mental picture' that the informant has of himself 'in his role as a smoker'. Changes in smoking habits are far more frequent than is generally thought to be the case, and so any information about them which refers to the past, based, as it must be, upon a general and personal assessment of current practices, is very likely to be in error ... recall is frequently mistaken both as regards the amount and the kind of smoking carried on. Apart from the special topical interest of this study, it has wide methodological implications which ought to be considered by all users of questionnaires.
From *Nature* 4 April 1959.

100 YEARS AGO

The influence of breed on egg-production in poultry is well seen in a report recently issued by Messrs. E. and W. Brown from University College, Reading. Danish, American, and English Leghorns were kept under comparable conditions for twelve months, and careful record was kept of the number of eggs laid. The Danish birds had been bred to yield a large number of eggs of moderate size; the English birds, on the other hand, had been largely bred for exhibition purposes, for which egg-producing capacity is not needed ... The profit on the English birds is shown to be much less than that on the Danish or American birds.
From *Nature* 1 April 1909.

50 & 100 YEARS AGO


W. AESCHBACH-HERTIG

Figure 1 | Bubbling up. The Wallender Born or 'Brubbel', a CO_2-driven cold-water geyser in the village of Wallenborn in western Germany, provides a natural illustration of CO_2 leakage from geological storage. Although largely harmless, such leakage would be undesirable in carbon-sequestration projects.

gas and carbon isotope data from nine natural gas fields in the United States, China and Hungary, the present study stands out by virtue of the large range of gas fields included and the methods used to identify the fate of the CO_2.

A central parameter of this analysis is the $CO_2/^3He$ ratio of the gases. The basic idea is that ^3He, a noble-gas isotope originating almost exclusively from Earth's mantle, behaves as a conservative tracer in the crustal environment of the gas reservoirs studied. The primary gas emplaced in these reservoirs has a characteristic $CO_2/^3He$ ratio, often indicating that it is of magmatic origin. Any reduction of this ratio is ascribed to the removal of CO_2 from the gas phase.

Gilfillan and colleagues' first, intriguing, finding is that declining $CO_2/^3He$ ratios in the gases are related to increasing concentrations of ^4He and ^{20}Ne. These correlations hold within individual fields as well as across the combined data set. The authors argue that this systematic behaviour strongly suggests that the gas has interacted with water, which provides a plausible source of crustal ^4He and atmospheric ^{20}Ne. Whereas the highly soluble CO_2 dissolves in the groundwater, the low-solubility noble gases He and Ne degas from the water into the gas phase, thereby producing the observed relationships. This indicates that solubility trapping is an important process, but does not rule out the possibility that mineral trapping also occurs.

A quantitative assessment of the contributions of the two trapping mechanisms is provided by a second line of evidence based on the $^{13}C/^{12}C$ isotope ratios of the CO_2 gas. This ratio is expected to change if CO_2 is removed by the formation of carbonate minerals, as the heavier isotope ^{13}C precipitates preferentially. Such an isotope fractionation also occurs as CO_2 dissolves in water, but to a lesser degree, depending on the prevailing pH conditions. By comparing the observed relationships between the $CO_2/^3He$ ratio (as a measure of CO_2 removal) and the $^{13}C/^{12}C$ isotope ratio in the different gas fields with models of the expected fractionation for either process, the authors show that the data are incompatible with mineral trapping, but can be explained by dissolution in water.

Gilfillan and colleagues' overall conclusion[1] is that in the nine gas fields investigated, covering different geological settings, solubility trapping played a major part, removing up to 90% or more of the initially emplaced CO_2. Mineral trapping played a minor part at best. Although dissolution in groundwater implies the possibility of CO_2 transport and eventual leakage to the atmosphere, as illustrated by Figure 1 and as is thought to occur in natural gas fields[4], this result does not mean that safe geological storage is impossible. But it highlights the need for a thorough assessment of the hydrogeological setting of prospective storage sites. And it demonstrates the power of the methods involved in assessing the effectiveness of different geochemical trapping mechanisms. ■

Werner Aeschbach-Hertig is at the Institut für Umweltphysik, Universität Heidelberg, D-69120 Heidelberg, Germany.
e-mail: aeschbach@iup.uni-heidelberg.de

1. Gilfillan, S. M. V. *et al. Nature* **458,** 614–618 (2009).
2. Metz, B. *et al.* (eds) *IPCC Special Report on Carbon Dioxide Capture and Storage* (Cambridge Univ. Press, 2005).
3. Schrag, D. P. *Science* **315,** 812–813 (2007).
4. Moore, J. *et al. Chem. Geol.* **217,** 365–385 (2005).

HIV

Immune memory downloaded

Dennis R. Burton and Pascal Poignard

An impressive system for retrieving large numbers of antibodies from memory B cells has been developed. It has been put into practice in an investigation of immune responses to the human immunodeficiency virus.

Infection of an individual with a virus or a bacterium triggers a vigorous response in white blood cells, some of which — B cells — are stimulated to produce antibodies that target the invading pathogen. The antibodies may be produced too late to prevent symptoms of infection, but the next contact with the same pathogen will probably be symptom-free as antibodies are rapidly deployed to clear the pathogen.

This antibody 'memory', which is crucial to vaccine efficacy, has two forms: antibodies circulating in the blood, made by a very-long-lived type of B cell in the bone marrow known as a plasma cell; and B cells in the blood that can be stimulated to make antibodies on contact with a pathogen[1,2]. The latter 'B-cell memory' carries a record of the antibodies an individual has made in response to a given pathogen, and is of great interest, not least in guiding the design of better vaccines. On page 636 of this issue, Scheid *et al.*[3] describe the detailed characterization of B-cell memory responses in the context of infection with the human immunodeficiency virus. The paper contains insights that are both of a general nature and likely to be specific to HIV.

Dissection of the B-cell memory response in human blood requires individual monoclonal antibodies (specific for particular sites on pathogen molecules) to be isolated from each B cell or each set (clone) of identical B cells. Scheid *et al.* accomplished this tour de force by selecting single-memory B cells specific for a preparation of the surface glycoproteins of HIV, amplifying antibody genes from each cell and then producing each antibody in a cell line (Fig. 1). In principle, sufficient numbers of B cells were sampled to reflect the full response to the glycoproteins. These glycoproteins were chosen because they are the sole target of antibodies able to neutralize the virus and prevent infection. The authors studied six HIV-infected donors whose blood sera can neutralize, to varying degrees, a range of different isolates of HIV. By analysing the antibody responses of the donors in detail, it was hoped to understand the origins of this broad neutralization.

Scheid and colleagues did most of their work on four donors, on average isolating more than 100 monoclonal antibodies to the surface glycoprotein preparation per donor. Each antibody was exhaustively characterized at the genetic and protein levels. The antibodies from each donor could be classified into 20–50 families of antibody, with varying numbers of close relatives in each family. The sequences of the antibodies in each family are highly divergent from the sequences characteristically found

in antibodies before contact with a pathogen, providing evidence that they are highly evolved to specifically recognize HIV glycoprotein. Constant exposure of the individuals' immune system to HIV over long periods is likely to be a significant factor here. Evolution is further reflected in the high affinities of the isolated antibodies for glycoprotein. Antibodies were found that bind across the whole surface of the glycoproteins, including to sites that have not been described previously.

Scheid *et al.* attempted to understand the neutralizing activities of the donor sera against a range of HIV isolates in terms of the activities of individual antibodies and combinations of antibodies. They were only partly successful. No single broadly neutralizing monoclonal antibodies were identified, so pools of monoclonal antibodies were tested. The pools for two donors showed neutralizing activity against representative HIV isolates, but only at high concentrations.

So it seems that further neutralizing antibodies remain unidentified in the donors. There are various possible reasons for the failure to find them — a potential disconnect between antibodies made by memory B cells and serum antibodies made by plasma cells in the bone marrow[4]; dysfunction of the memory B-cell compartment in HIV-infected individuals[5]; and Scheid and colleagues' use of a glycoprotein 'bait' that may inefficiently select memory B cells making neutralizing antibodies. The design of an optimal bait, which should ideally exactly mimic the conformation of glycoproteins on the surface of HIV, and indeed should thereby be a good vaccine candidate, is a recurring problem in this field.

One additional consideration that might help in understanding broad neutralization, using the approach of Scheid *et al.*, is the increasing access to HIV-infected individuals with exceptional broadly neutralizing serum activity[6]. The identification of broadly neutralizing monoclonal antibodies that target a sizeable proportion of the huge diversity of global HIV is highly desirable, as this will favour vaccine design[7]. Four such antibodies are already known to exist and, thanks to novel methods like that of Scheid *et al.*, new ones are certain to be forthcoming. The alternative possibility of a great number of antibodies, each targeting only a few HIV variants, is a less attractive basis for producing a practical vaccine.

In most of the donors studied by Scheid *et al.*, the HIV infection is under control. In some, the virus is kept to such low levels that the individuals concerned are known as elite controllers. But we must stress that there is no convincing evidence that antibody responses are responsible for the favourable clinical course seen in some HIV-infected people[8,9]. In contrast, however, there is strong evidence that neutralizing antibodies can prevent infection with HIV if those antibodies are present before exposure to the virus[10].

The work of Scheid and colleagues is an advance in attempts to clone human antibody responses. It will be interesting to see how the responses identified by this method compare with those obtained from other approaches and sources, such as 'gene rescue' from plasma cells of recently vaccinated individuals[11], and from large repertoires or libraries of immune and naive antibodies displayed on the surface of selectable particles such as phage[12]. It will, of course, also be essential to take any new-won understanding of protective antibody responses at the molecular level and exploit it in designing better vaccines[13]. ■

Dennis R. Burton and Pascal Poignard are in the Department of Immunology and Microbial Science, and the IAVI Neutralizing Antibody Center, The Scripps Research Institute, La Jolla, California 92037, USA.
e-mails: burton@scripps.edu; poignard@scripps.edu

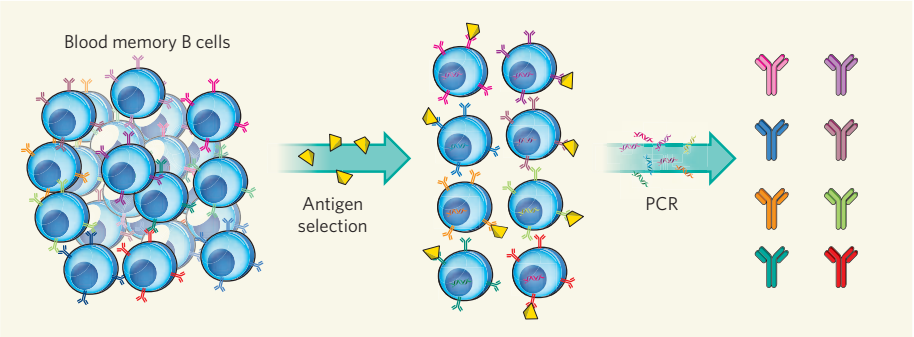


Figure 1 | Retrieving a B-cell memory response: Scheid and colleagues' approach[3]. The memory B-cell population is purified from the complex mixture of cells in blood by using specific cell-surface markers. Each memory B cell expresses on its surface multiple copies of a single antibody (Y-shapes). Each antibody can bind to a defined site on a given protein. The use of the protein (yellow) as a 'bait' allows the selection of all the memory B cells that make antibodies able to recognize that particular protein. In principle, any protein from any pathogen could be used for selection to interrogate an individual's memory B-cell response. Single B cells are then subjected to amplification of their antibody genes using the polymerase chain reaction (PCR), those genes then being incorporated into a cell line for producing antibodies. The end result is a large set of cell lines making monoclonal antibodies that can be individually characterized.

1. Wrammert, J. & Ahmed, R. *Biol. Chem.* **389,** 537–539 (2008).
2. Dörner, T. & Radbruch, A. *Immunity* **27,** 384–392 (2007).
3. Scheid, J. F. *et al. Nature* **458,** 636–640 (2009).
4. Guan, Y. *et al. Proc. Natl Acad. Sci. USA* **106,** 3952–3957 (2009).
5. Moir, S. *et al. J. Exp. Med.* **205,** 1797–1805 (2008).
6. Stamatatos, L., Morris, L., Burton, D. R. & Mascola, J. R. *Nature Med.* (in the press).
7. Karlsson Hedestam, G. B. *et al. Nature Rev. Microbiol.* **6,** 143–155 (2008).
8. Pereyra, F. *et al. J. Infect. Dis.* **197,** 563–571 (2008).
9. Bailey, J. R. *et al. J. Virol.* **80,** 4758–4770 (2006).
10. Mascola, J. R. *Curr. Mol. Med.* **3,** 209–216 (2003).
11. Wrammert, J. *et al. Nature* **453,** 667–671 (2008).
12. Lerner, R. A. *Angew. Chem. Int. Edn* **45,** 8106–8125 (2006).
13. Burton, D. R. *Nature Rev. Immunol.* **2,** 706–713 (2002).

NEUROSCIENCE

AMPA receptors get 'pickled'

Alexander C. Jackson and Roger A. Nicoll

In mediating fast synaptic communication in the brain, AMPA receptors require TARP auxiliary proteins. It seems that another distinct class of proteins also bind to AMPA receptors and regulate their function.

It is now well established that ion channels are not solitary creatures, but often have an entourage of auxiliary proteins. Indeed, voltage-gated potassium, sodium and calcium channels form stable complexes with an assortment of both cytoplasmic and transmembrane proteins that profoundly affect their localization and function[1]. The ligand-gated cation channels referred to as AMPA receptors (AMPARs) — a subtype of receptors activated by the neurotransmitter glutamate — are also known to robustly and selectively interact with a family of proteins termed transmembrane AMPA-receptor regulatory proteins (TARPs). As the first known examples of auxiliary subunits for ligand-gated ion channels, TARPs regulate both the surface expression and biophysical properties of AMPARs[2,3]. Writing in *Science*, Schwenk *et al.*[4] describe the unexpected interaction between AMPARs and another family of transmembrane proteins, named after the French word for a type of pickle — the cornichons. They find that, like TARPs, cornichons seem to influence both the intracellular trafficking and gating activity of AMPARs (Fig. 1, overleaf).

The regulation of AMPARs at excitatory

synapses between neurons are of particular interest, because plastic changes in the localization and function of these receptors are thought to underlie certain forms of learning and memory[5,6]. Stargazin, the prototypical TARP, was originally identified as being essential for the surface expression of AMPARs and for targeting them to synapses in granule cells of the cerebellum. Apart from stargazin, which is also called γ-2, the TARP family is now known to include γ-3, γ-4, γ-5, γ-7 and γ-8. These transmembrane proteins are widely expressed in the central nervous system and are intimately involved with AMPARs throughout their lives — from synthesis to surface expression and synaptic targeting[2,3].

TARP proteins localize to synapses through motifs in their carboxy terminus that bind to the PDZ domain of scaffolding proteins, such as PSD-95, in postsynaptic neurons[2,3]. TARPs are also powerful modulators of AMPAR gating and pharmacology: they slow channel deactivation and desensitization; enhance single-channel conductance; convert the partial agonist kainate into a full agonist; and cause the competitive antagonist CNQX to act as a partial agonist[7].

Schwenk and colleagues' data[4], however, indicate that TARPs are not the only intimates in AMPARs' inner circle. The authors used a proteomic approach to uncover the identity of proteins in the rat brain that interact with AMPAR subunits. They detected two proteins that had not previously been linked with glutamate-receptor trafficking or synaptic transmission — CNIH-2 and CNIH-3.

These members of the mammalian CNIH family are homologous to the cornichon proteins, which have been characterized primarily in flies and yeast. In both the fruitfly *Drosophila* and mammals, cornichon is a cargo receptor necessary for the export of epidermal growth factor receptor (EGFR) ligands from the endoplasmic reticulum, a subcellular organelle[8,9]. This common mechanism of action underscores the remarkable phylogenetic conservation of function among cornichon proteins[10]. In this context, a close association with AMPARs seems to be a decidedly extracurricular activity for the cornichons.

Schwenk *et al.* posit that a surprisingly small proportion (30%) of AMPARs associate with TARPs, with the remaining 70% forming complexes with CNIHs. The proportion of TARP-associated AMPARs proposed may be an underestimate, however, as the authors used an antibody directed against γ-2/3 as a proxy for all TARPs. In fact, the other TARPs, including γ-4, γ-5, γ-7 and γ-8, are also expressed in the brain and exhibit a robust association with AMPARs[2,3,11,12].

Nevertheless, the suggestion that native AMPARs can be parsed into mutually exclusive pools — one associated with TARPs and another with CNIHs — is intriguing. As TARPs have carboxy-terminus PDZ-binding motifs and CNIHs do not, it is tempting to speculate that there is a division of labour between these two sets of auxiliary proteins in their handling of AMPARs. Are there two trafficking pathways for AMPARs, one TARP-dependent and the other CNIH-dependent? Are TARPs and CNIHs interchanged during their transport from one subcellular compartment to another, or from extrasynaptic sites to synaptic sites? Can TARPs, CNIHs and AMPARs form ternary complexes? And could it be that a portion of the CNIH-associated pool of AMPARs remains in the endoplasmic reticulum, reflecting the established role of CNIHs in trafficking EGFR ligands?

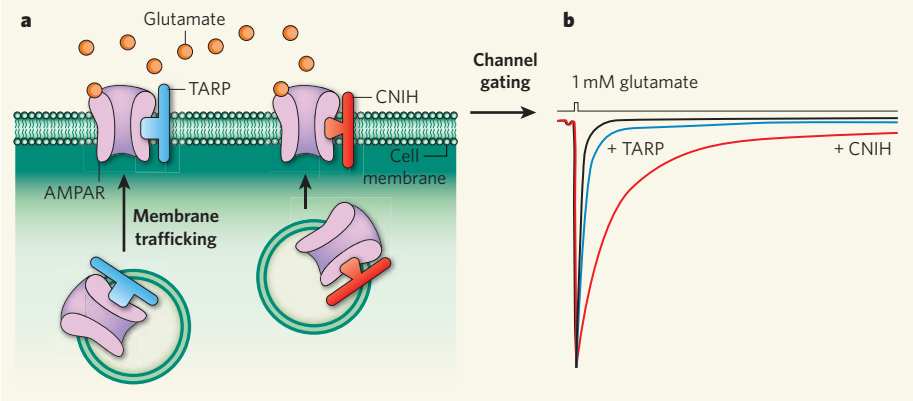


Figure 1 | AMPA receptors expand their circle of friends. **a**, Schwenk *et al.*[4] show that, in addition to TARPs, the AMPA subtype of glutamate receptors (AMPARs) can bind to another group of transmembrane proteins — CNIHs. Like TARPs, CNIHs mediate trafficking of AMPARs to the cell surface. **b**, Moreover, CNIHs slow the deactivation (illustrated) and desensitization of AMPARs that have been activated by glutamate. (**b** adapted from ref. 4.)

Apart from forming complexes with AMPAR subunits, CNIH-2 and CNIH-3 share other features with TARPs. Like TARPs, CNIHs are widely distributed in the brain and are expressed in principal neurons, interneurons and glial cells in the brain's hippocampus, cerebellum and neocortex. A clear exception is cerebellar granule cells, in which CNIHs are conspicuously absent and in which surface expression and synaptic targeting of AMPARs have been shown to rely on γ-2 (refs 2, 3). It is also interesting to note that two other members of the mammalian cornichon family, CNIH-1 and CNIH-4, are widely expressed in the mouse brain[13], although to date they have no clear neuronal function. Whether the differential expression of TARPs and CNIHs is cell-type specific, and how their functions segregate or overlap in single cells, are questions that are likely to pique the curiosity of researchers in the field.

Another property that CNIHs share with TARPs is that they not only modulate AMPAR trafficking, but also dramatically slow the deactivation (Fig. 1b) and desensitization kinetics of these receptors, thus potentially enhancing the charge transfer associated with synaptic events[2–4,7]. Intriguingly, the magnitude of CNIHs' effect on AMPAR kinetics greatly outstrips that of γ-2. It will be interesting to assess what other effects CNIHs have on the biophysical properties and pharmacology of AMPARs, especially when compared with the established effects of TARPs. For instance, what is the influence of CNIH association on the single-channel conductance of AMPARs? Do CNIHs dramatically influence kainate efficacy, like TARPs? Is glutamate affinity altered by CNIH–AMPAR interactions? And can CNIHs modify TARP-associated AMPARs, or vice versa?

These are exhilarating times for the study of glutamate-receptor regulation. Several other candidate auxiliary subunits for ionotropic glutamate receptors have also emerged in the past few years: NETO1 and NETO2 for kainate receptors[14], NETO1 for NMDA receptors[15], and SOL-1 for GLR-1 receptors in the nematode worm *Caenorhabditis elegans*[16]. Along with cornichons, these discoveries add richness and diversity, as well as further complexity, to our view of glutamate-receptor regulation in the nervous system. Having identified these new players, it will be of great interest to investigate their potential roles in development, in synaptic-plasticity mechanisms associated with learning and memory, and in the mechanisms underlying disease. ■

Alexander C. Jackson and Roger A. Nicoll are in the Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, California 94143, USA.
e-mail: nicoll@cmp.ucsf.edu

1. Vacher, H. *et al. Physiol. Rev.* **88,** 1407–1447 (2008).
2. Nicoll, R. A. *et al. Science* **311,** 1253–1256 (2006).
3. Ziff, E. B. *Neuron* **53,** 627–633 (2007).
4. Schwenk, J. *et al. Science* **323,** 1313–1319 (2009).
5. Malinow, R. & Malenka, R. C. *Annu. Rev. Neurosci.* **25,** 103–126 (2002).
6. Bredt, D. S. & Nicoll, R. A. *Neuron* **40,** 361–379 (2003).
7. Milstein, A. D. & Nicoll, R. A. *Trends Pharmacol. Sci.* **29,** 333–339 (2008).
8. Roth, S. *et al. Cell* **81,** 967–978 (1995).
9. Bökel, C. *et al. Development* **133,** 459–470 (2006).
10. Castro, C. P. *et al. J. Cell Sci.* **120,** 2454–2466 (2007).
11. Kato, A. S. *et al. Neuron* **59,** 986–996 (2008).
12. Soto, D. *et al. Nature Neurosci.* **12,** 277–285 (2009).
13. Lein, E. S. *et al. Nature* **445,** 168–176 (2007).
14. Zhang, W. *et al. Neuron* **61,** 385–396 (2009).
15. Ng, D. *et al. PLoS Biol.* **7,** e41 (2009).
16. Zheng, Y. *et al. Nature* **427,** 451–457 (2004).

Q&A

Deep view — a slice of the Hubble Space Telescope's view of the visible Universe.

COSMOLOGY

Dark matter and dark energy

Robert Caldwell and Marc Kamionkowski

Observations continue to indicate that the Universe is dominated by invisible components — dark matter and dark energy. Shedding light on this cosmic darkness is a priority for astronomers and physicists.

What is the composition of the Universe?

In terms of their contribution to the mean energy density, the contents of the Universe are approximately 75% dark energy, 20% dark matter and 5% normal (atomic) matter, with smaller contributions from photons and neutrinos. These measurements rely on the validity of the hot Big Bang model, general relativity and the cosmological principle (that the Universe is uniform on the largest scales). The breadth and depth of experiments and observations that support these underlying tenets give us confidence that this model of the cosmos has a solid foundation.

What is the evidence for dark matter?

We can infer the presence of dark matter through indirect methods, despite not being able to see it (Fig. 1, overleaf). Newton's laws state that the mass of a body can be determined by the motion of its satellites. Thus, it has been calculated that the mass of galaxy clusters is far larger than that of their constituent galaxies, and that the mass of galaxies is far larger than the combined mass of their constituent stars and interstellar gas. And there is plenty more corroborating evidence. Yet there is very good reason to expect that this extra 'stuff' is not normal matter. Such an abundance of normal matter would be difficult to conceal from the prying eyes of astronomers, and would furthermore leave a distinct signature in the cosmic microwave background (CMB) radiation (relic radiation from the Big Bang), and in the properties of galaxies and clusters, that is simply not seen.

Why can't we conclude that Newton's laws break down at the distance scales of galaxies or clusters?

This might have been a reasonable hypothesis a few decades ago. However, any alternative gravity theory that accounts for the observed galaxy and cluster dynamics must also explain the vast body of data on gravitational lensing (the deflection of light from distant sources), the CMB and large-scale structures. At the same time, it must also satisfy a suite of precise constraints on gravity obtained within the Solar System.

How much dark matter is there nearby?

The orbital velocities of stars in the Milky Way suggest a mean mass density of dark matter in our neighbourhood of about a third of a proton mass per cubic centimetre. For perspective, this is 10^6 times greater than the mean density of the cosmos, but 24 orders of magnitude smaller than the mean density of water. Because whatever objects make up dark matter move in the same Galactic gravitational potential well as stars, we know that they must be moving with velocities of about 200 kilometres per second. Earth's orbit around the Sun implies that the amount of dark matter incident on the Earth varies by about 10% from summer to winter (Fig. 1). Furthermore, the distribution of galactic dark matter may not be smooth; galaxy formation is an ongoing process, and computational studies suggest that there may be a significant amount of dark-matter substructure in the form of clumps and tidal streams.

What is the best bet for the nature of dark matter?

From the vast array of proposals, the most promising ideas involve novel elementary particles. Among the candidates that have withstood long-standing theoretical scrutiny are weakly interacting massive particles (WIMPs) and axions. WIMPs, like neutrinos, interact only weakly with ordinary matter. They arise naturally in extensions to the standard model of particle physics (for example, in supersymmetry or in models with large extra dimensions). Detection of WIMPs is one of the primary goals of the Large Hadron Collider (LHC) at CERN near Geneva, Switzerland. The other candidate, the axion, is an elementary particle hypothesized to explain some of the symmetries of the strong interactions that bind quarks in protons and neutrons. There are other possibilities, so it is necessary to keep an open mind. However, constraints on the strength of the interaction of dark-matter particles with ordinary matter, their stability against decay and their 'coldness' — dark-matter particles today must move slowly compared with the speed of light — allow the range of possibilities to be pared down.

What experiments or observations can help?

Clearly, the most compelling resolution to the dark-matter problem would be the direct detection of dark-matter particles. Currently,

there are some 20 experimental projects seeking to detect WIMPs by observing the 10–100 kiloelectronvolts of energy that would be deposited in a detector when a WIMP from the Galactic halo scatters from an atomic nucleus in the detector and makes it recoil. The target nuclei in some of these experiments are located in metallic crystals; the nuclear recoil is then detected through the recoiling energy collected in the detector. The challenge in these and other dark-matter detection experiments is to distinguish the signature of dark matter from the crowd of terrestrial-radiation backgrounds. But the current generation of experiments is becoming sufficiently sensitive that it will soon be possible to vet some of the leading particle-physics models for dark matter. The discovery of unknown particles at the LHC would greatly narrow the range of dark-matter candidates and boost our confidence that we are on the right track. But it would not eliminate the need for an *in situ* astrophysical detection.

Haven't there already been claims of dark-matter detection?

Yes. The DAMA experiment, operating deep underground at the Gran Sasso National Laboratory in Italy, has reported detection of the tell-tale annual modulation in dark-matter flux consistent with Earth's orbit through the Galactic dark-matter halo. This signal has not been corroborated by other experiments. Because other experiments use different target nuclei, the various results can only be compared in the context of specific theories of dark matter. The mass of the simplest, 'supersymmetric' WIMPs and their couplings to normal matter, proposed to explain the DAMA result, have been excluded by the other experiments.

How else can we see dark matter?

Although individual WIMPs are in theory stable, pairs of WIMPs can 'annihilate', producing high-energy photons and cosmic rays in the form of positrons (antielectrons), antiprotons and neutrinos. Detection of such particles might provide indirect evidence for dark matter. The most likely nearby sources of these annihilation products would be the Galactic Centre, where the dark-matter density is high, or the cores of some of the dark-matter-dominated dwarf galaxies surrounding the Milky Way (Fig. 1). One telling clue would be monoenergetic γ-rays. There is a host of ground-based, balloon- and satellite-borne experiments looking for these clues.

What about the cosmic-ray experiments ...

In 2008, PAMELA (Payload for Antimatter Matter Exploration and Light-nuclei Astrophysics), a satellite-borne cosmic-ray experiment, and the balloon-borne ATIC (Advanced Thin Ionization Calorimeter) experiment, reported an excess flux of high-energy cosmic-ray positrons. These observations might be a consequence of WIMP annihilation, but the observed flux is higher, by several orders of magnitude, than the simplest WIMP models predict. One interpretation is that WIMP dark matter is more complicated than previously thought. However, more prosaic astrophysical explanations (such as particle acceleration by nearby pulsars) must be excluded before the anomaly can be attributed to dark matter.

... and future possibilities for studying dark matter?

Experiments to detect dark matter directly

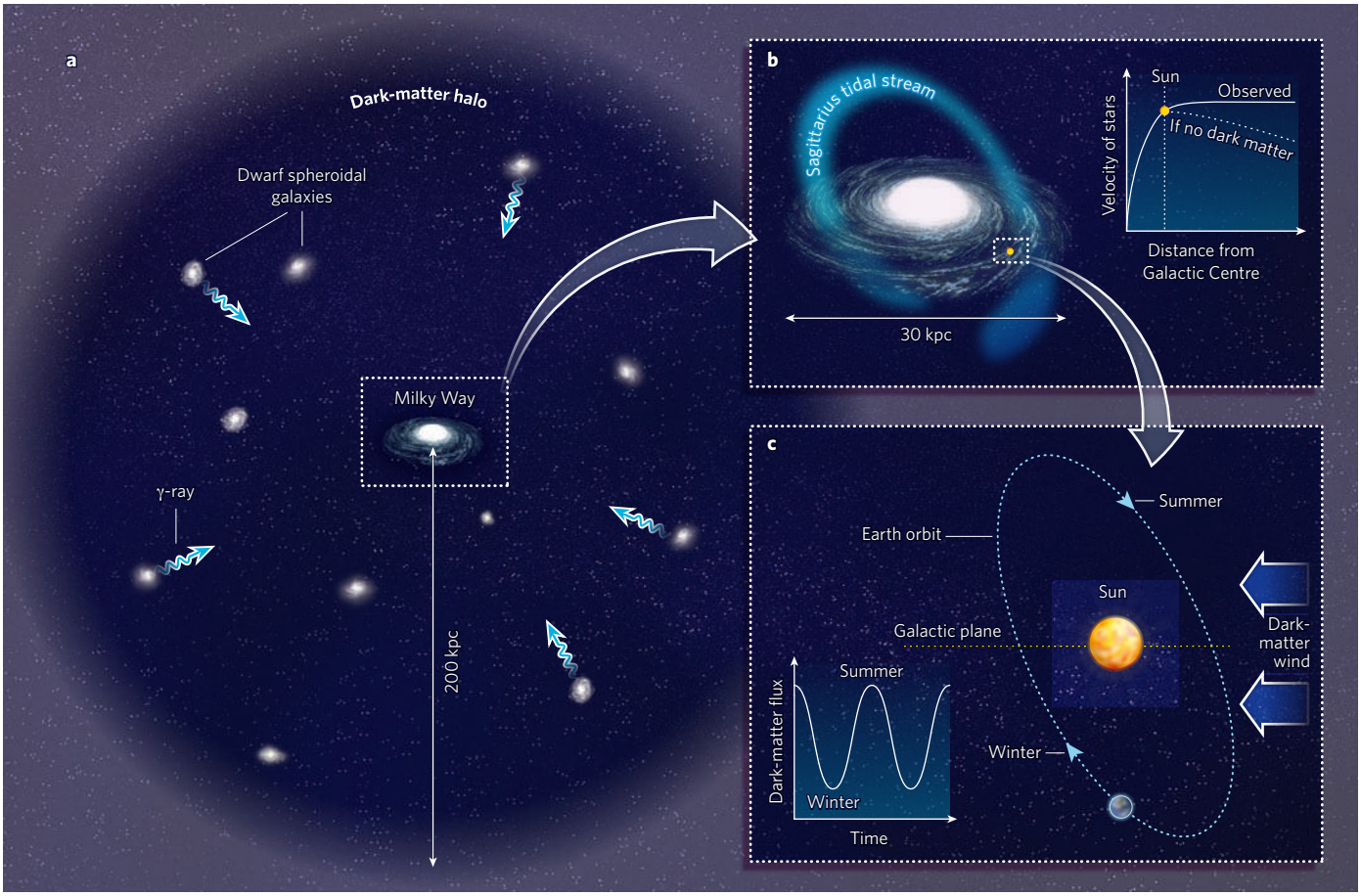


Figure 1 | Dark matter and how it might be detected. a, b, The rotational velocity of its stars and gas indicates that the Milky Way is embedded in a dark-matter halo extending out to a radius of about 200 kiloparsecs (kpc). High-energy γ-rays may be produced by the annihilation of dark-matter particles in neighbouring dwarf spheroidal galaxies and near the Galactic Centre, where the dark-matter density is expected to be highest. The dark-matter density may also be enhanced in the tidal stream of matter that trails from the Sagittarius dwarf galaxy and entangles the Milky Way. **c,** Earth's orbit through the Galactic dark-matter halo may produce a modulation of the dark-matter flux identified in experiments that aim to detect dark matter directly: a smaller (by about 10%) flux is expected when Earth moves in the same direction as the dark-matter 'wind' from the Galactic halo (in winter) than when it moves against it (in summer).

aim to exploit the dark-matter WIMP 'wind' (Fig. 1) and isolate the characteristic annual modulation in the WIMP flux from other background signals of terrestrial origin. Meanwhile, Gaia, a satellite mission set to launch in the near future, aims to chart the position and motion of about 10^9 nearby stars; this map will be used to trace out the gravitational field of the Milky Way, and thereby infer the dark-matter distribution in its dark-matter halo. A variety of experiments, including that using the recently launched Fermi Gamma-ray Space Telescope, will look for γ-rays from WIMP annihilation. And high-energy neutrino telescopes, such as IceCube at the South Pole, will look for neutrinos produced by the annihilation of WIMPs that have accumulated in the Sun and Earth.

What about dark energy?

The observation that the expansion of the Universe is speeding up (Fig. 2), instead of slowing down owing to the mutual gravitational attraction of matter, indicates that there is much more to the Universe than we understand at present. The leading interpretation is that the Universe is filled by something — dubbed dark energy — that 'antigravitates'. Whereas the possibility for gravitational repulsion does not exist in Newtonian gravity, it does exist in general relativity. The equivalence between matter and energy means that gaseous pressures caused by thermal molecular motions can be a source of gravitational fields. The gravitational field of a fluid with sufficiently negative pressure is repulsive. Although it may be difficult to imagine how molecular motions can give rise to a negative pressure, it has been realized that some of the quantum fields that arise in elementary-particle theory allow for fluids with negative pressure. Dark energy is thus simply the negative-pressure fluid that is postulated to account for cosmic acceleration.

What is the best bet for the nature of dark energy?

The simplest candidate for dark energy is Einstein's cosmological constant, which denotes a perfectly uniform fluid with negative pressure that is associated with the lowest energy (vacuum) state of the Universe. However, the observationally required value of the cosmological constant is 10^{120} times smaller than the theoretical expectation. Alternatively, dark energy might be due to a fluid of unknown particles, similar to the axion but much smaller in mass — quantum theory predicts that such particles could supply the requisite negative pressure to accelerate the cosmic expansion.

How reliable are the known laws of gravitation on cosmological scales?

General relativity works. It has been extremely well tested in the Solar System, and it is used to make sense of a vast catalogue of astrophysical and cosmological observations. These successes do not preclude the possibility of variations in the laws of gravitation on cosmological length scales. A Pandora's box of gravitational theories has been proposed to explain the accelerated cosmic expansion. But it is proving surprisingly difficult to tinker with gravity without running up against the precision constraints in the Solar System, and so far there are no compelling alternatives.

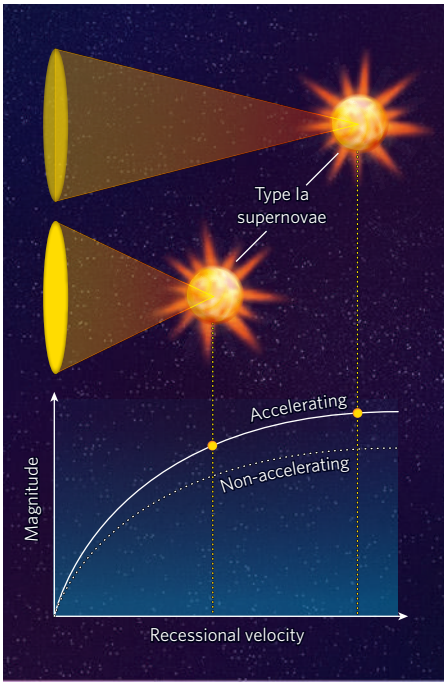


Figure 2 | Cosmic acceleration and dark energy. Type Ia supernovae, which result from the explosion of white-dwarf stars, are thought to be standard candles (objects of known brightness). This property allows astronomers to determine how far away such supernovae are, based on their apparent brightness as observed on Earth — the dimmer the object seems to be, the higher the value of its magnitude and the farther away it is. The observation that these supernovae are dimmer than expected, at a given recessional velocity, has led to the conclusion that the Universe's expansion has been accelerating over approximately the past 5 billion years, before which the expansion was decelerating. The cause of this cosmic acceleration is widely attributed to dark energy.

Could dark matter and dark energy be related?

It seems reasonable to consider the possibility of a 'dark sector', beyond the standard model of particle physics, containing a dark-matter particle and a dark-energy field. Both seem to require unknown sources of gravitational fields, one attractive and the other repulsive, but there have been no convincing proposals that unify the two phenomena.

Could cosmic acceleration be caused by any other phenomena?

One might consider new forms of gravitation (whereby normal matter produces the same antigravitational effect as dark energy), new electromagnetic effects (whereby distant supernovae are artificially dimmed; Fig. 2), or some other flaw in our fundamental assumptions (such as the statistical homogeneity and isotropy of the Universe on the largest length scales). The current state of observations does not favour one of these alternatives, but we must keep an open mind.

What recent observations have helped to refine the dark-energy problem?

The observations of 'baryon acoustic oscillations' have been used to corroborate and refine the evidence for cosmic acceleration. These cosmic ripples made by primordial sound waves are imprinted on the CMB and on the distribution of galaxies. By measuring how the wavelength of the ripples varies with the distance from Earth, one can chart the history of the cosmic expansion.

What experiments can help to determine the nature of dark energy?

There is a decided absence of compelling theoretical explanations for the physics underlying cosmic acceleration, so the approach to date has been to gather more of the same type of data in the hope that some clue will pop out. Apart from using supernovae and baryon acoustic oscillations, other methods that measure the rate at which normal and dark matter cluster under the influence of gravitation in an accelerating Universe are also progressing. One promising technique uses gravitational lensing in the 'weak' regime to set constraints on dark energy; in this regime, instead of the strong bending of light that results in highly distorted images in the form of elongated arcs, the images of distant sources are only weakly stretched and magnified by foreground matter (the 'lenses'). Another technique uses X-ray emissions of hot gas in galaxy clusters to determine the depth of their gravitational potential wells. But despite the promise of these methods, it may be difficult to determine the underlying physics of cosmic acceleration. On the other hand, this seems to be the only way to tackle such a challenging and fundamental problem. ■

Robert Caldwell is in the Department of Physics and Astronomy, Dartmouth College, Hanover, New Hampshire 03755, USA. Marc Kamionkowski is at the California Institute of Technology, Pasadena, California 91125, USA.
e-mails: robert.r.caldwell@dartmouth.edu; kamion@tapir.caltech.edu

FURTHER READING

Caldwell, R. & Kamionkowski, M. The physics of cosmic acceleration. *Annu. Rev. Nucl. Part. Sci.* (in the press); preprint available at http://arxiv.org/abs/0903.0866 (2009).
Frieman, J. A., Turner, M. S. & Huterer, D. Dark energy and the accelerating Universe. *Annu. Rev. Astron. Astrophys.* **46,** 385–432 (2008).
Hooper, D. & Baltz, E. A. Strategies for determining the nature of dark matter. *Annu. Rev. Nucl. Part. Sci.* **58,** 293–314 (2008).
Hogan, J. & Brumfiel, G. Unseen Universe. *Nature* **448,** 240–248 (2007).

Is there an association between NPY and neuroticism?

Arising from: Z. Zhou *et al. Nature* 452, 997–1001 (2008)

Psychiatric genetics has been hampered by the fact that initially exciting findings from underpowered studies are so often not replicated in larger, more powerful, data sets. Here we show that the claims of Zhou *et al.*[1] that neuropeptide Y (*NPY*) diplotype-predicted expression is correlated with trait anxiety (neuroticism) is not replicated in a data set consisting of phenotypically extreme individuals drawn from a large ($n = 88{,}142$) non-clinical population. We found no association between *NPY* diplotype or diplotype-predicted expression and neuroticism. Our reply to Zhou and colleagues forms part of a larger debate[2–5] (see, for example, http://www.nature.com/news/2008/080709/full/454154a.html) about the efficacy and replicability of candidate driven versus genome wide approaches to psychiatric genetics.

In their recent study, Zhou and colleagues[1] used a candidate gene driven approach to select NPY for investigation as a possible modulator of genetic susceptibility to anxiety and neuroticism. Zhou *et al.* concluded that "haplotype-driven *NPY* expression…inversely correlates with trait anxiety" and that their results "help to explain inter-individual variation in resiliency to stress, a risk factor for many diseases"[1].

To test their claims we genotyped all seven single nucleotide polymorphisms (SNPs) investigated by Zhou *et al.*[1] in 582 singletons from the extreme 5% tails of the Eysenck Personality Questionnaire neuroticism score distribution from a non-clinical population of 88,142 individuals from the south-west of England[2]. This sample has close to 100% power to detect a genetic effect accounting for 1.25% of phenotypic variance at an alpha level of 0.01. As Zhou *et al.* state that NPY explains between 3.3% and 3.4% of variance in trait anxiety[1], we have close to 100% power to test their claims.

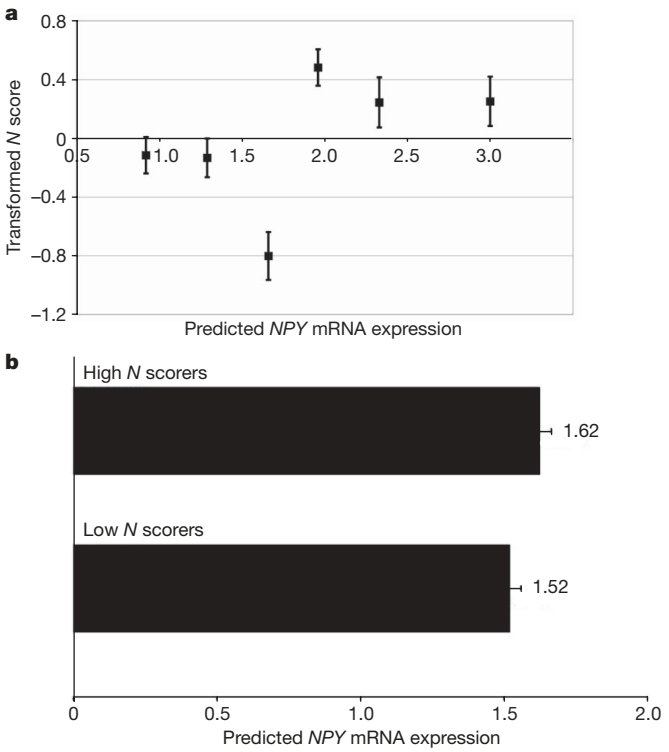


Figure 1 | Diplotype-predicted *NPY* expression and neuroticism. **a**, Regression of transformed age and sex-regressed *N* scores (mean ± s.e.m.) and diplotype-predicted expression values in 507 subjects (from left to right: H1/H1, $n = 151$; H1/H3, $n = 129$; H3/H3, $n = 16$; H1/H2, $n = 139$; H2/H3, $n = 33$; and H2/H2, $n = 39$). **b**, Diplotype-predicted *NPY* mRNA expression levels (mean and s.e.m.) of high neuroticism scorers ($n = 265$) and low neuroticism scorers ($n = 242$) compared with a two-tailed *t*-test ($P = 0.06$).

Diplotypes were assigned to each sample using the five haplotype definitions outlined by Zhou and colleagues[1]. The three most common haplotypes (H1, H2 and H3) formed six common diplotypes that had each been assigned an expression profile on the basis of lymphoblast *NPY* messenger RNA levels: low (LL:H1/H1), intermediate (LH:H1/H3, H3/H3 and H1/H2) and high (HH:H2/H3 and H2/H2). Subjects with minor diplotypes ($n = 75$) were not included in further analyses. Figure 1a shows the distribution of neuroticism scores by diplotype-predicted mRNA expression levels. Neuroticism was compared among diplotype groups by analysis of variance (ANOVA) and regression analysis. The diplotype-predicted values of mRNA expression were taken from Zhou *et al.*[1] as predicted by a co-dominant model. One-way ANOVA on all samples demonstrated no effect of *NPY* diplotype on neuroticism phenotype ($F(5) = 1.38$; $P = 0.14$) nor of *NPY*-diplotype-predicted expression ($F(2) = 1.01$; $P = 0.36$). Furthermore, *NPY*-diplotype-predicted expression was not correlated with transformed age and sex-regressed neuroticism scores (Fig. 1a). Furthermore, *NPY* diplotype-predicted mRNA levels did not differ significantly between subjects with high and low neuroticism scores ($P = 0.06$; Fig. 1b).

If *NPY* diplotype does in fact exert an effect on neuroticism, then the main effect size must be smaller than 1.25% and probably smaller than 0.5% (power = 87.6%). This lack of replication highlights the problems inherent in candidate gene driven approaches to psychiatric genetics.

METHODS

Oligonucleotide primers specific for seven different SNP markers (rs3037354, rs17149106, rs16147, rs16139, rs9785023, rs5574 and rs16475) were used to amplify the target *NPY* fragments by PCR. Sequencing was performed with Sequenom's MassARRAY technology[6].

Statistical power was calculated by simulation methods and implemented in Perl[2]. We ran 1,000 simulations of effect sizes ranging from 2.0% to 0.1% and using either 0.05 or 0.01 alpha levels, and calculated the proportion of times that a significant result was obtained.

Colleen H. Cotton[1], Jonathan Flint[1] & Thomas G. Campbell[1,2]
[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.
[2]St. Cross College, University of Oxford, Oxford OX1 3LZ, UK.
e-mail: thomasgordoncampbell@gmail.com

Received 11 November 2008; accepted 11 February 2009.

1. Zhou, Z. *et al.* Genetic variation in human *NPY* expression affects stress response and emotion. *Nature* **452**, 997–1001 (2008).
2. Willis-Owen, S. A. *et al.* The serotonin transporter length polymorphism, neuroticism, and depression: a comprehensive assessment of association. *Biol. Psychiatry* **58**, 451–456 (2005).
3. Munafo, M. R., Bowes, L., Clark, T. G. & Flint, J. Lack of association of the *COMT* ($Val^{158/108}$ Met) gene and schizophrenia: a meta-analysis of case-control studies. *Mol. Psychiatry* **10**, 765–770 (2005).
4. Munafo, M. R. *et al.* Genetic polymorphisms and personality in healthy adults: a systematic review and meta-analysis. *Mol. Psychiatry* **8**, 471–484 (2003).
5. Abbott, A. Psychiatric genetics: The brains of the family. *Nature* **454**, 154–157 (2008).
6. Gabriel, S. & Ziaugra, L. SNP genotyping using Sequenom MassARRAY 7K platform. *Curr. Protoc. Hum. Genet.* Chapter 2, Unit 2.12 (2004).

doi:10.1038/nature07927

Zhou et al. *reply*

Replying to: C. H. Cotton, J. Flint & T. G. Campbell *Nature* **458,** doi:10.1038/nature07927 (2009)

The inability of Cotton *et al.*[1] to detect an effect of a functional haplotype (and locus) of neuropeptide Y (NPY), a stress regulatory neuropeptide, on neuroticism is interesting. Although it is important to measure effects of functional loci on complex behaviours, the strength of our study[2], and primary basis of its conclusions, was the larger and convergent effects of NPY on intermediate phenotypes, including regional brain responses to emotional stimuli and pain, and brain *NPY* messenger RNA and plasma NPY levels. Eysenck Neuroticism is a trait that we did not directly investigate. We reported modest association of NPY with two Harm Avoidance subscales from the Tridimensional Personality Questionnaire. Association of NPY with the complex trait of anxiety, especially when measured differently, is not the first place we would look to validate our results.

Concerning their advocacy of genome-wide approaches, if we follow the conclusions of their genome-wide association study with the same data set[3] then no loci contribute >1% of the variance in neuroticism. This is plausible, and could explain why they found no effect of NPY. However, Cotton *et al.*[1] genotyped the extremes of a large but relatively uncharacterized sample. Theoretically powerful, this approach may in practice be problematic. At the extremes of the distribution various confounds such as severe environmental stresses, rare functional alleles and measurement errors are more likely to be over-represented. Their study did not identify new functional loci for anxiety nor confirm functional loci for which there is independent evidence, as mentioned later. It is reasonable to request evidence that a tool works before using it to 'weed the garden'.

There is indeed debate as to how to proceed in gene discovery for behaviour. However, candidate gene and genome-wide approaches are not at war. The goal of genome-wide studies is to identify locations of functional polymorphisms. Studies using intermediate phenotypes, on which alleles exert larger effects than complex behaviours, may be better able to expand our understanding of mechanism. Consistent and convergent effects of several functional alleles on intermediate phenotypes have demonstrated the validity of this approach. Recent discoveries relating common alleles to behaviour have primarily relied on brain imaging tools. Examples include the serotonin-transporter-linked polymorphic region (5-HTTLPR) that has a weak effect on depression and anxiety—an association that was indeed obscured when only the extremes of the distribution were compared[4]—but strong effects on brain metabolic responses to emotional stimuli[5] and the uncoupling of limbic feedback circuitry (accounting for 30% of the variance in anxious temperament[6]). Brain imaging studies have also shown that a functional missense variant (Val158Met) of COMT alters brain activity during cognition[7], pain[8] and response to emotional stimuli (accounting for 38% of the variance in emotionality[9]), while having much more modest effects on complex behaviours, including anxiety. If allele effects on crudely measured behavioural phenotypes are undetectable in very large data sets, this may suggest that genome-wide genetic methods should be applied to data sets of more modest size, in which intermediate phenotypes have been measured that are more robust in detecting genetic influences on behaviour.

Zhifeng Zhou[1], Guanshan Zhu[1]†, Ahmad R. Hariri[2], Mary-Anne Enoch[1], David Scott[3], Rajita Sinha[4], Matti Virkkunen[5], Deborah C. Mash[6], Robert H. Lipsky[1], Xian-Zhang Hu[1], Colin A. Hodgkinson[1], Ke Xu[1], Beata Buzas[1], Qiaoping Yuan[1], Pei-Hong Shen[1], Robert E. Ferrell[2], Stephen B. Manuck[2], Sarah M. Brown[2], Richard L. Hauger[7], Christian S. Stohler[8], Jon-Kar Zubieta[3] & David Goldman[1]

[1]Laboratory of Neurogenetics, NIAAA, NIH, Bethesda, Maryland 20892, USA.
e-mail: davidgoldman@mail.nih.gov
[2]Departments of Psychiatry, Human Genetics, and Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA.
[3]Departments of Psychiatry and Radiology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA.
[4]Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut 06510, USA.
[5]Department of Psychiatry, University of Helsinki, Helsinki 00014, Finland.
[6]Department of Neurology, University of Miami School of Medicine, Miami, Florida 33124, USA.
[7]Department of Psychiatry, San Diego VA Healthcare System and University of California, San Diego, California 92161, USA.
[8]School of Dentistry, University of Maryland, Baltimore, Maryland 21201, USA.
†Present address: Innovation Centre China, AstraZeneca Global R&D, Shanghai 201203, China.

1. Cotton, C. H., Flint, J. & Campbell, T. G. Is there an association between NPY and neuroticism? *Nature* **458,** doi:10.1038/nature07927 (2009).
2. Zhou, Z. *et al.* Genetic variation in human *NPY* expression affects stress response and emotion. *Nature* **452,** 997–1001 (2008).
3. Shifman, S. *et al.* A whole genome association study of neuroticism using DNA pooling. *Mol. Psychiatry* **13,** 302–312 (2008).
4. Sirota, L. A., Greenberg, B. D., Murphy, D. L. & Hamer, D. H. Non-linear association between the serotonin transporter promoter polymorphism and neuroticism: a caution against using extreme samples to identify quantitative trait loci. *Psychiatr. Genet.* **9,** 35–38 (1999).
5. Hariri, A. R. *et al.* Serotonin transporter genetic variation and the response of the human amygdala. *Science* **297,** 400–403 (2002).
6. Pezawas, L. *et al.* 5-HTTLPR polymorphism impacts human cingulate-amygdala interactions: a genetic susceptibility mechanism for depression. *Nature Neurosci.* **8,** 828–834 (2005).
7. Egan, M. F. *et al.* Effect of COMT $Val^{108/158}$ Met genotype on frontal lobe function and risk for schizophrenia. *Proc. Natl Acad. Sci. USA* **98,** 6917–6922 (2001).
8. Zubieta, J.-K. *et al.* COMT $val^{158}met$ genotype affects μ-opioid neurotransmitter responses to a pain stressor. *Science* **299,** 1240–1243 (2003).
9. Smolka, M. N. *et al.* Catechol-*O*-methyltransferase *valmet* genotype affects processing of emotional stimuli in the amygdala and prefrontal cortex. *J. Neurosci.* **25,** 836–842 (2005).

doi:10.1038/nature07928

Is there an association between NPY and neuroticism?

Arising from: Z. Zhou *et al. Nature* 452, 997–1001 (2008)

Psychiatric genetics has been hampered by the fact that initially exciting findings from underpowered studies are so often not replicated in larger, more powerful, data sets. Here we show that the claims of Zhou *et al.*[1] that neuropeptide Y (*NPY*) diplotype-predicted expression is correlated with trait anxiety (neuroticism) is not replicated in a data set consisting of phenotypically extreme individuals drawn from a large ($n = 88{,}142$) non-clinical population. We found no association between *NPY* diplotype or diplotype-predicted expression and neuroticism. Our reply to Zhou and colleagues forms part of a larger debate[2–5] (see, for example, http://www.nature.com/news/2008/080709/full/454154a.html) about the efficacy and replicability of candidate driven versus genome wide approaches to psychiatric genetics.

In their recent study, Zhou and colleagues[1] used a candidate gene driven approach to select NPY for investigation as a possible modulator of genetic susceptibility to anxiety and neuroticism. Zhou *et al.* concluded that "haplotype-driven *NPY* expression…inversely correlates with trait anxiety" and that their results "help to explain inter-individual variation in resiliency to stress, a risk factor for many diseases"[1].

To test their claims we genotyped all seven single nucleotide polymorphisms (SNPs) investigated by Zhou *et al.*[1] in 582 singletons from the extreme 5% tails of the Eysenck Personality Questionnaire neuroticism score distribution from a non-clinical population of 88,142 individuals from the south-west of England[2]. This sample has close to 100% power to detect a genetic effect accounting for 1.25% of phenotypic variance at an alpha level of 0.01. As Zhou *et al.* state that NPY explains between 3.3% and 3.4% of variance in trait anxiety[1], we have close to 100% power to test their claims.

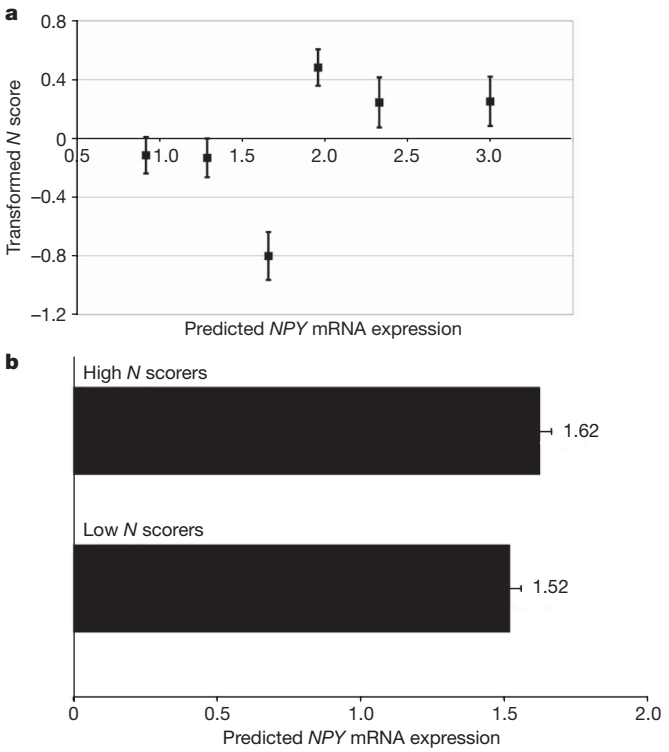


Figure 1 | Diplotype-predicted *NPY* expression and neuroticism. **a**, Regression of transformed age and sex-regressed *N* scores (mean ± s.e.m.) and diplotype-predicted expression values in 507 subjects (from left to right: H1/H1, $n = 151$; H1/H3, $n = 129$; H3/H3, $n = 16$; H1/H2, $n = 139$; H2/H3, $n = 33$; and H2/H2, $n = 39$). **b**, Diplotype-predicted *NPY* mRNA expression levels (mean and s.e.m.) of high neuroticism scorers ($n = 265$) and low neuroticism scorers ($n = 242$) compared with a two-tailed *t*-test ($P = 0.06$).

Diplotypes were assigned to each sample using the five haplotype definitions outlined by Zhou and colleagues[1]. The three most common haplotypes (H1, H2 and H3) formed six common diplotypes that had each been assigned an expression profile on the basis of lymphoblast *NPY* messenger RNA levels: low (LL:H1/H1), intermediate (LH:H1/H3, H3/H3 and H1/H2) and high (HH:H2/H3 and H2/H2). Subjects with minor diplotypes ($n = 75$) were not included in further analyses. Figure 1a shows the distribution of neuroticism scores by diplotype-predicted mRNA expression levels. Neuroticism was compared among diplotype groups by analysis of variance (ANOVA) and regression analysis. The diplotype-predicted values of mRNA expression were taken from Zhou *et al.*[1] as predicted by a co-dominant model. One-way ANOVA on all samples demonstrated no effect of *NPY* diplotype on neuroticism phenotype ($F(5) = 1.38$; $P = 0.14$) nor of *NPY*-diplotype-predicted expression ($F(2) = 1.01$; $P = 0.36$). Furthermore, *NPY*-diplotype-predicted expression was not correlated with transformed age and sex-regressed neuroticism scores (Fig. 1a). Furthermore, *NPY* diplotype-predicted mRNA levels did not differ significantly between subjects with high and low neuroticism scores ($P = 0.06$; Fig. 1b).

If *NPY* diplotype does in fact exert an effect on neuroticism, then the main effect size must be smaller than 1.25% and probably smaller than 0.5% (power = 87.6%). This lack of replication highlights the problems inherent in candidate gene driven approaches to psychiatric genetics.

METHODS

Oligonucleotide primers specific for seven different SNP markers (rs3037354, rs17149106, rs16147, rs16139, rs9785023, rs5574 and rs16475) were used to amplify the target *NPY* fragments by PCR. Sequencing was performed with Sequenom's MassARRAY technology[6].

Statistical power was calculated by simulation methods and implemented in Perl[2]. We ran 1,000 simulations of effect sizes ranging from 2.0% to 0.1% and using either 0.05 or 0.01 alpha levels, and calculated the proportion of times that a significant result was obtained.

Colleen H. Cotton[1], Jonathan Flint[1] & Thomas G. Campbell[1,2]
[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.
[2]St. Cross College, University of Oxford, Oxford OX1 3LZ, UK.
e-mail: thomasgordoncampbell@gmail.com

Received 11 November 2008; accepted 11 February 2009.

1. Zhou, Z. *et al.* Genetic variation in human *NPY* expression affects stress response and emotion. *Nature* **452**, 997–1001 (2008).
2. Willis-Owen, S. A. *et al.* The serotonin transporter length polymorphism, neuroticism, and depression: a comprehensive assessment of association. *Biol. Psychiatry* **58**, 451–456 (2005).
3. Munafo, M. R., Bowes, L., Clark, T. G. & Flint, J. Lack of association of the *COMT* ($Val^{158/108}$ Met) gene and schizophrenia: a meta-analysis of case-control studies. *Mol. Psychiatry* **10**, 765–770 (2005).
4. Munafo, M. R. *et al.* Genetic polymorphisms and personality in healthy adults: a systematic review and meta-analysis. *Mol. Psychiatry* **8**, 471–484 (2003).
5. Abbott, A. Psychiatric genetics: The brains of the family. *Nature* **454**, 154–157 (2008).
6. Gabriel, S. & Ziaugra, L. SNP genotyping using Sequenom MassARRAY 7K platform. *Curr. Protoc. Hum. Genet.* Chapter 2, Unit 2.12 (2004).

doi:10.1038/nature07927

Zhou et al. *reply*

Replying to: C. H. Cotton, J. Flint & T. G. Campbell *Nature* **458,** doi:10.1038/nature07927 (2009)

The inability of Cotton *et al.*[1] to detect an effect of a functional haplotype (and locus) of neuropeptide Y (NPY), a stress regulatory neuropeptide, on neuroticism is interesting. Although it is important to measure effects of functional loci on complex behaviours, the strength of our study[2], and primary basis of its conclusions, was the larger and convergent effects of NPY on intermediate phenotypes, including regional brain responses to emotional stimuli and pain, and brain *NPY* messenger RNA and plasma NPY levels. Eysenck Neuroticism is a trait that we did not directly investigate. We reported modest association of NPY with two Harm Avoidance subscales from the Tridimensional Personality Questionnaire. Association of NPY with the complex trait of anxiety, especially when measured differently, is not the first place we would look to validate our results.

Concerning their advocacy of genome-wide approaches, if we follow the conclusions of their genome-wide association study with the same data set[3] then no loci contribute >1% of the variance in neuroticism. This is plausible, and could explain why they found no effect of NPY. However, Cotton *et al.*[1] genotyped the extremes of a large but relatively uncharacterized sample. Theoretically powerful, this approach may in practice be problematic. At the extremes of the distribution various confounds such as severe environmental stresses, rare functional alleles and measurement errors are more likely to be over-represented. Their study did not identify new functional loci for anxiety nor confirm functional loci for which there is independent evidence, as mentioned later. It is reasonable to request evidence that a tool works before using it to 'weed the garden'.

There is indeed debate as to how to proceed in gene discovery for behaviour. However, candidate gene and genome-wide approaches are not at war. The goal of genome-wide studies is to identify locations of functional polymorphisms. Studies using intermediate phenotypes, on which alleles exert larger effects than complex behaviours, may be better able to expand our understanding of mechanism. Consistent and convergent effects of several functional alleles on intermediate phenotypes have demonstrated the validity of this approach. Recent discoveries relating common alleles to behaviour have primarily relied on brain imaging tools. Examples include the serotonin-transporter-linked polymorphic region (5-HTTLPR) that has a weak effect on depression and anxiety—an association that was indeed obscured when only the extremes of the distribution were compared[4]—but strong effects on brain metabolic responses to emotional stimuli[5] and the uncoupling of limbic feedback circuitry (accounting for 30% of the variance in anxious temperament[6]). Brain imaging studies have also shown that a functional missense variant (Val158Met) of COMT alters brain activity during cognition[7], pain[8] and response to emotional stimuli (accounting for 38% of the variance in emotionality[9]), while having much more modest effects on complex behaviours, including anxiety. If allele effects on crudely measured behavioural phenotypes are undetectable in very large data sets, this may suggest that genome-wide genetic methods should be applied to data sets of more modest size, in which intermediate phenotypes have been measured that are more robust in detecting genetic influences on behaviour.

Zhifeng Zhou[1], Guanshan Zhu[1]†, Ahmad R. Hariri[2], Mary-Anne Enoch[1], David Scott[3], Rajita Sinha[4], Matti Virkkunen[5], Deborah C. Mash[6], Robert H. Lipsky[1], Xian-Zhang Hu[1], Colin A. Hodgkinson[1], Ke Xu[1], Beata Buzas[1], Qiaoping Yuan[1], Pei-Hong Shen[1], Robert E. Ferrell[2], Stephen B. Manuck[2], Sarah M. Brown[2], Richard L. Hauger[7], Christian S. Stohler[8], Jon-Kar Zubieta[3] & David Goldman[1]

[1]Laboratory of Neurogenetics, NIAAA, NIH, Bethesda, Maryland 20892, USA.
e-mail: davidgoldman@mail.nih.gov
[2]Departments of Psychiatry, Human Genetics, and Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA.
[3]Departments of Psychiatry and Radiology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA.
[4]Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut 06510, USA.
[5]Department of Psychiatry, University of Helsinki, Helsinki 00014, Finland.
[6]Department of Neurology, University of Miami School of Medicine, Miami, Florida 33124, USA.
[7]Department of Psychiatry, San Diego VA Healthcare System and University of California, San Diego, California 92161, USA.
[8]School of Dentistry, University of Maryland, Baltimore, Maryland 21201, USA.
†Present address: Innovation Centre China, AstraZeneca Global R&D, Shanghai 201203, China.

1. Cotton, C. H., Flint, J. & Campbell, T. G. Is there an association between NPY and neuroticism? *Nature* **458,** doi:10.1038/nature07927 (2009).
2. Zhou, Z. *et al.* Genetic variation in human *NPY* expression affects stress response and emotion. *Nature* **452,** 997–1001 (2008).
3. Shifman, S. *et al.* A whole genome association study of neuroticism using DNA pooling. *Mol. Psychiatry* **13,** 302–312 (2008).
4. Sirota, L. A., Greenberg, B. D., Murphy, D. L. & Hamer, D. H. Non-linear association between the serotonin transporter promoter polymorphism and neuroticism: a caution against using extreme samples to identify quantitative trait loci. *Psychiatr. Genet.* **9,** 35–38 (1999).
5. Hariri, A. R. *et al.* Serotonin transporter genetic variation and the response of the human amygdala. *Science* **297,** 400–403 (2002).
6. Pezawas, L. *et al.* 5-HTTLPR polymorphism impacts human cingulate-amygdala interactions: a genetic susceptibility mechanism for depression. *Nature Neurosci.* **8,** 828–834 (2005).
7. Egan, M. F. *et al.* Effect of COMT $Val^{108/158}$ Met genotype on frontal lobe function and risk for schizophrenia. *Proc. Natl Acad. Sci. USA* **98,** 6917–6922 (2001).
8. Zubieta, J.-K. *et al.* COMT $val^{158}met$ genotype affects μ-opioid neurotransmitter responses to a pain stressor. *Science* **299,** 1240–1243 (2003).
9. Smolka, M. N. *et al.* Catechol-*O*-methyltransferase *valmet* genotype affects processing of emotional stimuli in the amygdala and prefrontal cortex. *J. Neurosci.* **25,** 836–842 (2005).

doi:10.1038/nature07928

ARTICLES

Tyrosine dephosphorylation of H2AX modulates apoptosis and survival decisions

Peter J. Cook[1,2]*, Bong Gun Ju[1,3]*, Francesca Telese[1], Xiangting Wang[1], Christopher K. Glass[4] & Michael G. Rosenfeld[1]

Life and death fate decisions allow cells to avoid massive apoptotic death in response to genotoxic stress. Although the regulatory mechanisms and signalling pathways controlling DNA repair and apoptosis are well characterized, the precise molecular strategies that determine the ultimate choice of DNA repair and survival or apoptotic cell death remain incompletely understood. Here we report that a protein tyrosine phosphatase, EYA, is involved in promoting efficient DNA repair rather than apoptosis in response to genotoxic stress in mammalian embryonic kidney cells by executing a damage-signal-dependent dephosphorylation of an H2AX carboxy-terminal tyrosine phosphate (Y142). This post-translational modification determines the relative recruitment of either DNA repair or pro-apoptotic factors to the tail of serine phosphorylated histone H2AX (γ-H2AX) and allows it to function as an active determinant of repair/survival versus apoptotic responses to DNA damage, revealing an additional phosphorylation-dependent mechanism that modulates survival/apoptotic decisions during mammalian organogenesis.

The developmentally regulated transcriptional cofactor EYA is a component of the retinal determination pathway that controls the development of various organ systems in metazoans, including the kidney[1–3]. The primary phenotypic consequence of loss of EYA activity is increased apoptotic cell death in early tissue primordium and subsequent agenesis of target tissues[3–5]. Previous work by our laboratory and others identified a phosphatase enzymatic domain in mammalian EYA1–4 as well as the *Drosophila* homologue *eyes absent* (*eya*), and demonstrated that EYA is a functional phosphatase[6–8]. Although early *in vitro* phosphatase assays using synthetic phosphopeptides indicated that EYA might possess dual specificity, subsequent data have indicated that, *in vivo*, EYA primarily functions as a tyrosine phosphatase[9]. Here, we demonstrate that increased apoptosis seen in the absence of EYA is at least in part due to persistent phosphorylation of H2AX Y142, a mark that is a component of the mechanisms that distinguish between apoptotic and repair responses to genotoxic stress.

EYA–H2AX interactions

We noticed that increased apoptosis and loss of renal tubules seen in the developing kidney of *Eya1*$^{-/-}$ mouse embryos coincided with increased immunostaining for serine-139-phosphorylated H2AX (γ-H2AX) (Fig. 1a, b and Supplementary Fig. 1). Nuclear phosphorylation of the histone variant H2AX was recently shown to be a crucial component of apoptosis induced by the activation of the JNK/SAPK stress response pathway[10], in addition to having a well studied role in DNA damage repair[11–14]. Because the developing kidney is exposed to localized hypoxia during early development as the rapidly proliferating organ outgrows the local vasculature, potentially leading to activation of stress response pathways and increased generation of reactive oxygen species[15,16], we considered the possibility that apoptosis induced in the absence of EYA might be related to altered DNA-damage-response pathways. To mimic the events in the *Eya1*$^{-/-}$ kidney in a cell model, we depleted endogenous EYA1 or EYA3 in 293T human embryonic kidney cells using specific short interfering RNAs (siRNAs; Supplementary Fig. 2) and then subjected the cells to hypoxic conditions for 20 h. EYA1 and EYA3 have been previously qualified as phosphatase enzymes[6–8] and both are expressed in 293T cells. Notably, knockdown of either *EYA1* or *EYA3* using specific siRNAs caused a significant increase in TdT-mediated dUTP nick end labelling (TUNEL)-positive apoptotic nuclei in response to hypoxia (Fig. 1c). Analogous experiments directly inducing DNA damage with ionizing radiation resulted in a similar increase in sensitivity for EYA-depleted cells (Supplementary Fig. 3). Thus, in embryonic kidney cells, both *in vivo* and in culture, an increase in apoptotic cell death is observed in the absence of EYA1 that may be related to the cellular response to DNA damage, which involves γ-H2AX[11,17].

We therefore investigated a potential interaction between EYA and H2AX by co-immunoprecipitation assays using 293T embryonic kidney cells before and after exposing the cells to ionizing radiation to induce DNA damage. We could detect interactions between H2AX and wild-type EYA1 or EYA3 only under DNA damage conditions both using transfected, tagged expression constructs for EYA1/EYA3 and H2AX (Fig. 2a), and when examining endogenous EYA3 and H2AX proteins with specific antibodies (Fig. 2b). EYA was capable of interacting with H2AX in the context of chromatin, based on co-immunoprecipitation experiments using fixed sonicated chromatin from 293T cells as input (Fig. 2c). In response to ionizing-radiation-induced double-stranded DNA breaks, H2AX is phosphorylated by

[1]Howard Hughes Medical Institute School of Medicine, University of California, San Diego, California 92037, USA. [2]Department of Biology Graduate Program, School of Medicine, University of California, San Diego, California 92093, USA. [3]Department of Life Science, Sogang University, Seoul 121-742, Korea. [4]Department of Cellular and Molecular Medicine, School of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA.
*These authors contributed equally to this work.

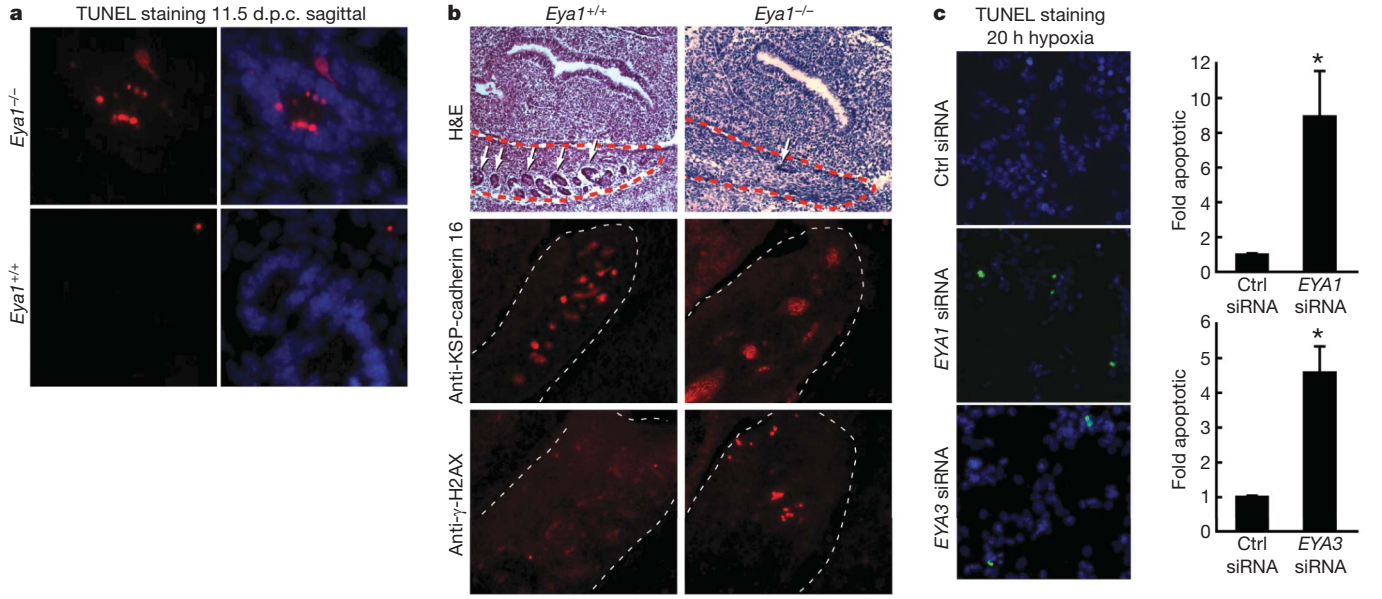


Figure 1 | Loss of EYA leads to increased γ-H2AX-posititve apoptotic cells. **a**, TUNEL staining reveals apoptotic cells within the developing kidney of *Eya1*$^{-/-}$ embryos at embryonic day (E)11.5 not present in wild-type littermate mice. d.p.c., days post coitum. Original magnification, ×20. **b**, Abnormal morphology and loss of developing renal tubules (white arrows) within the urogenital ridge (red dotted line) in *Eya1*$^{-/-}$ embryos coincides with increased γ-H2AX-positive nuclei by immunostaining. H&E, haematoxylin and eosin. Original magnification, ×5. **c**, In culture, 293T human embryonic kidney cells depleted for EYA1 and EYA3 using siRNA displayed increased apoptotic response to hypoxia for 20 h (2% O_2). Cell counts were performed on TUNEL-stained cells co-stained with 4,6-diamidino-2-phenylindole (DAPI) in triplicate to identify the proportion of TUNEL-positive nuclei. The basal level of apoptosis under these conditions was 1.4% TUNEL-positive/total nuclei. Bar graphs represent mean ± s.e.m. of fold apoptotic cells normalized to control siRNA from triplicate samples. Asterisk indicates $P < 0.05$. Original magnification, ×10.

ATM/ATR phosphatidylinositol-3-OH kinase (PI(3)K)-family kinases on chromatin, forming long stretches of serine-phosphorylated γ-H2AX flanking the break, visible as γ-H2AX immunostained foci[18]. Endogenous EYA3 co-immunoprecipitated γ-H2AX in 293T cells after ionizing radiation treatment (Fig. 2b, lower panel), and immunostaining of transfected haemagglutinin (HA)-tagged EYA1 or EYA3 protein in 293T embryonic kidney cells revealed a clear co-localization of EYA with γ-H2AX foci after treatment with ionizing radiation (Fig. 2d, e).

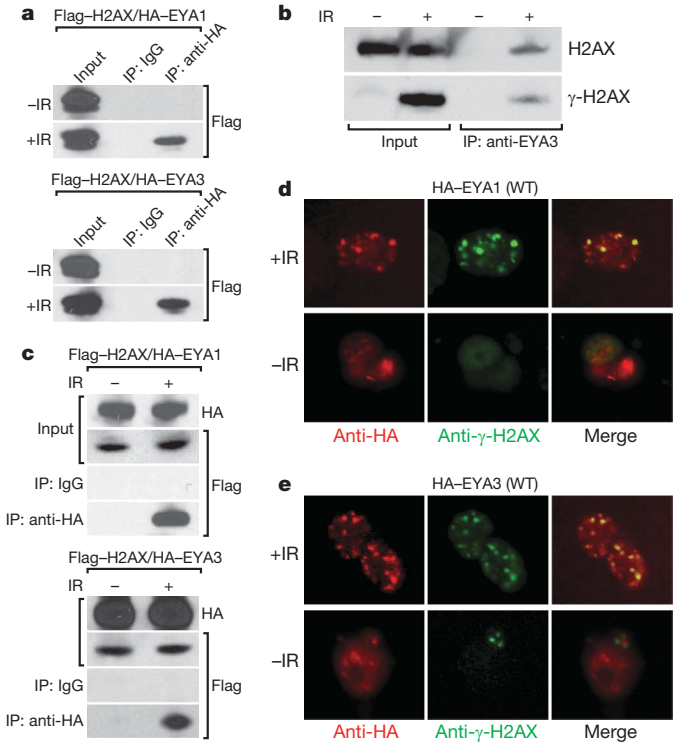


Figure 2 | EYA interacts with H2AX in a DNA-damage-dependent manner. **a**, HA-tagged EYA1 or EYA3 interacts with Flag-tagged H2AX in 293T cells in response to ionizing radiation (IR; 5 Gy), but not under basal conditions. **b**, Co-immunoprecipitation experiments examining endogenous EYA3 protein using a specific EYA3 antibody recapitulated that interaction data for the tagged proteins. **c**, Using sonicated chromatin as input, co-immunoprecipitation experiments showed that HA–EYA1/3 interacts with H2AX on chromatin. **d**, **e**, Immunostaining of 293T cells demonstrates that transfected, HA-tagged EYA1 (**d**) or EYA3 (**e**) localizes to DNA-damage-induced foci coincident with γ-H2AX specifically after treatment with ionizing radiation (5 Gy, 1 h). Representative examples of foci formation are shown. Original magnification, ×40.

These results suggest that in response to damage, EYA is recruited to H2AX foci that mark DNA double-strand breaks. To test this formally, we used the oestrogen receptor-I PpoI system[19,20], in which 4-hydroxytamoxifen (4-OHT) is used to induce activation of the eukaryotic homing endonuclease I-PpoI that then generates double-stranded breaks at defined genomic loci, including a site on chromosome 1 within an intron of the *DAB1* locus. Chromatin immunoprecipitation analysis after 4-OHT induction of I-PpoI in 293T cells revealed that γ-H2AX and EYA3 were present at a 6 h time point at a 4-kilobase (kb) region flanking the I-PpoI cut site, which is consistent with a direct role for EYA in the cellular response to genotoxic stress (Supplementary Fig. 4).

Interestingly, we found that EYA3 is serine-phosphorylated in 293T cells in response to genotoxic stress (Fig. 3a), consistent with the recent identification of EYA3 as a potential substrate for the DNA-damage-response protein kinases ATM and ATR[21–23]. Inhibition of ATM/ATR function, by pre-treating cells with the PI(3)K inhibitor caffeine, blocked the interaction between EYA3 or EYA1 and H2AX in response to ionizing radiation (Fig. 3b). Serine 219 of EYA3 was identified by mass spectroscopy as a target residue for ATM/ATR phosphorylation[22], and a S219A EYA3 mutant failed to form damage-dependent nuclear foci or interact with H2AX after ionizing radiation treatment (Fig. 3c, d), indicating that ATM/ATR phosphorylation of EYA3 on serine 219 is crucial for directing EYA–H2AX interactions. Because

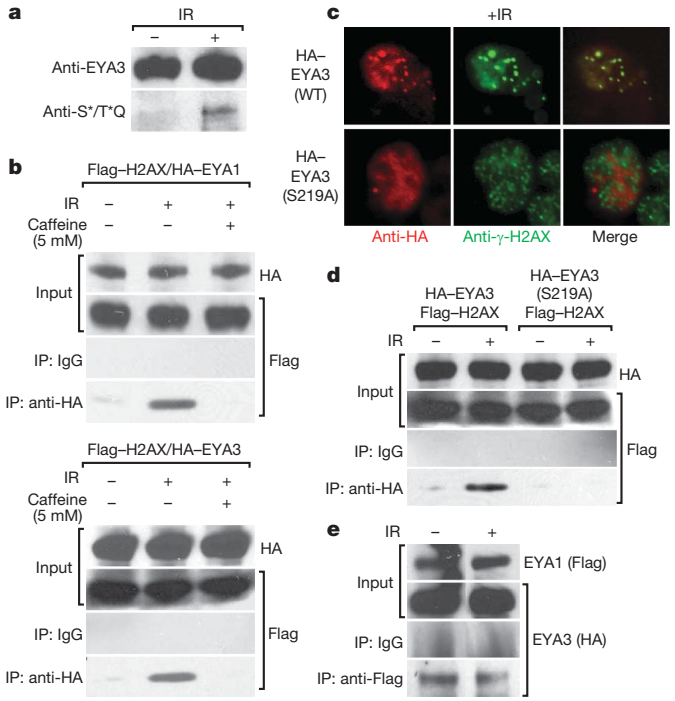


Figure 3 | EYA3 phosphorylation by ATM/ATR DNA-damage-dependent kinases regulates the interaction between EYA and H2AX. a, Endogenous EYA3 was immunoprecipitated from 293T cells with a specific EYA3 antibody and western blotting was performed with an antibody specific to the phosphorylated target site of ATM/ATR, demonstrating phosphorylation of EYA3 in response to DNA damage (5 Gy IR). **b**, EYA1/3 interaction with H2AX is lost in the presence of a PI(3)K inhibitor (5 mM caffeine). **c**, Mutation of the ATM/ATR phosphorylation site of EYA3 (S219) prevents formation of damage-induced EYA3 foci. Representative examples of foci formation are shown. Original magnification, ×40. **d**, HA–EYA3 (S219A) fails to interact with Flag–H2AX in response to DNA damage (5 Gy IR) by co-immunoprecipitation in 293T cells. **e**, DNA-damage-independent interaction of EYA3 and EYA1 was assessed by co-immunoprecipitation in 293T cells.

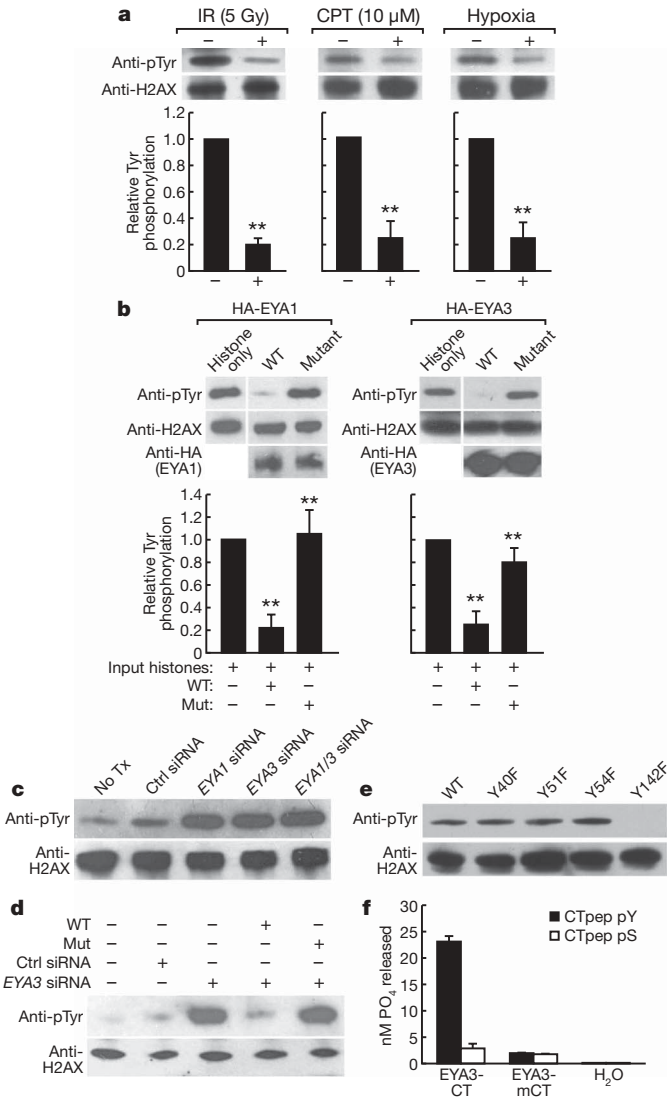


Figure 4 | Tyrosine phosphorylated H2AX is a substrate for EYA phosphatase. a, Immunoprecipitation western blot (IP-western) of tyrosine phosphorylated H2AX in response to DNA damage signals. Bars represent quantified western blot signals normalized to untreated cells. **b**, *In vitro* phosphatase assay using immunopurified wild-type EYA1/3 or enzymatically inactive mutant proteins (EYA1 D323A, EYA3 D246A) and bovine histone. Bars represent quantified western blot signals normalized to input. Mean values ± s.e.m. from triplicate western blot experiments are shown. Double asterisk indicates $P < 0.001$. **c**, siRNA knockdown of endogenous EYA1/3 in 293T cells (48 h) and subsequent IP-western for tyrosine phosphorylated H2AX. No Tx indicates non-transfected cells. **d**, Rescue of EYA function by co-transfection of human siRNA and murine wild-type or enzymatically inactive mutant EYA3 constructs in 293T human embryonic kidney cells reveals loss of H2AX phosphotyrosine mark dependent on EYA phosphatase activity. **e**, Individual substitution mutations of four H2AX tyrosine residues followed by IP-western to detect phosphotyrosine. **f**, *In vitro* phosphatase assay using bacterially expressed EYA3 EYA domain, wild type or D246A, with purified peptides of the H2AX tail (amino acids 128–142) phosphorylated at S139 (CTpep pS) or Y142 (CTpep pY) demonstrates that EYA phosphatase activity is specific for phosphotyrosine. The Michaelis constant (K_m) value for EYA dephosphorylation of CTpep pY was 0.38 mM with a corresponding K_{cat}/K_m value of 0.96 M^{-1} min^{-1}. Bar graphs represent mean ± s.e.m. of nM PO_4 released from triplicate phosphatase reactions.

EYA1 and EYA3 are seen to interact in 293T embryonic kidney cells both before and after treatment with ionizing radiation (Fig. 3e), we suspect that regulation of EYA3 via damage-dependent phosphorylation at serine 219 is one cue that may direct both EYA1 and EYA3 to γ-H2AX, indicating that these covalent modifications of H2AX and EYA may act as sensors for the DNA-damage-response pathway.

H2AX is an EYA tyrosine phosphatase substrate

We next tested whether the interaction between H2AX and EYA could represent a substrate–enzyme relationship. Because current evidence suggests that EYA is a tyrosine-specific phosphatase[6–8], we assessed its activity as a tyrosine phosphatase on γ-H2AX. H2AX purified either from 293T cells or from bovine histone fraction possesses tyrosine phosphorylation as seen using a phosphotyrosine-specific antibody (Supplementary Fig. 5). This tyrosine phosphorylation mark on H2AX decreased in response to DNA damage induced by ionizing radiation, the topoisomerase I inhibitor CPT, or hypoxia (Fig. 4a). To determine whether this H2AX phosphorylation mark might be a target of EYA phosphatase activity, we used an *in vitro* phosphatase assay, mixing immunopurified HA-tagged EYA1 or EYA3 with H2AX protein. Wild-type EYA effectively removed the phosphotyrosine mark from H2AX, whereas the phosphatase-inactive mutant EYA proteins (EYA1 D323A or EYA3 D246A) had little or no effect (Fig. 4b).

To confirm this activity in a cellular context, 293T human embryonic kidney cells were transfected with siRNA against *EYA1* or *EYA3* or control siRNA and subsequently exposed to ionizing radiation. In contrast to untransfected cells or cells receiving control siRNA, which displayed a loss of γ-H2AX tyrosine phosphorylation in response to damage as seen previously, EYA siRNA-treated cells showed significantly increased γ-H2AX tyrosine phosphorylation levels as assessed

by western blot analysis (Fig. 4c). Knockdown of EYA1 or EYA3 had no effect on tyrosine phosphorylation of H2AX in 293T cells not exposed to ionizing radiation (Supplementary Fig. 6). Rescuing EYA function by expressing wild-type murine EYA3 (Fig. 4d) or EYA1 (Supplementary Fig. 7) constructs—not targeted by the siRNAs—into these siRNA-depleted cells reversed this increased H2AX phosphorylation, whereas a phosphatase-dead mutant EYA failed to rescue EYA function. The observation that depletion of either EYA1 or EYA3 alone proved to be sufficient to block H2AX tyrosine dephosphorylation fully in these cells suggested a lack of compensatory activity by these two homologues. Because EYA1 and EYA3 co-purify in 293T cells before and after damage (Fig. 3e), we are tempted to suggest that, specifically in the context of this embryonic kidney cell line model, EYA1 and EYA3 may form a stable complex which exhibits tyrosine phosphatase activity towards γ-H2AX, with both components required for the overall stability of the enzymatic complex, although these factors may be non-redundant *in vivo*.

We next sought to identify precisely which tyrosine residue(s) on H2AX were phosphorylated. Mutagenesis of each of the four tyrosine residues in H2AX revealed that only mutation of tyrosine residue 142 blocked H2AX tyrosine phosphorylation as assessed by western blot analysis (Fig. 4e), indicating that Y142 was the only phosphorylated tyrosine.

To confirm the *in vitro* tyrosine phosphatase function of EYA[6–8] and demonstrate specificity for tyrosine-phosphorylated H2AX, rather than serine, we generated a bacterially expressed construct representing the enzymatically active C-terminal EYA domain of EYA3 (ref. 6). This EYA enzyme showed robust phosphatase activity when mixed with a synthetic phosphopeptide representing the C-terminal tail domain of H2AX (CT-pep) phosphorylated on tyrosine, but showed minimal activity towards a serine-phosphorylated tail peptide (Fig. 4f). These data biochemically establish the ability of EYA to directly dephosphorylate H2AX phosphorylated on Y142.

H2AX Y142 dephosphorylation: function in apoptosis

To begin to evaluate a possible connection between EYA-mediated tyrosine dephosphorylation of H2AX Y142 and modulation of the apoptotic response, we examined the function of this phosphotyrosine mark in the context of the DNA damage response. Flag-tagged H2AX Y142F mutant was phosphorylated on S139 in response to damage, although at levels significantly lower than Flag-tagged wild-type H2AX (Fig. 5a). Time course analysis of S139 phosphorylation of H2AX Y142F in response to 10 Gy ionizing radiation in 293T human embryonic kidney cells revealed consistently reduced levels compared to wild type between 1 and 8 h (Supplementary Fig. 8). Thus, whereas Y142 phosphorylation does not function as a prerequisite for S139 phosphorylation in DNA damage response[24], it may have a significant role in promoting or maintaining serine phosphorylation by DNA-damage-response kinases.

It has been established that a key function of H2AX S139 phosphorylation is to provide a docking site for DNA repair factors near or at DNA double-strand breaks[18]. These factors include mediator of

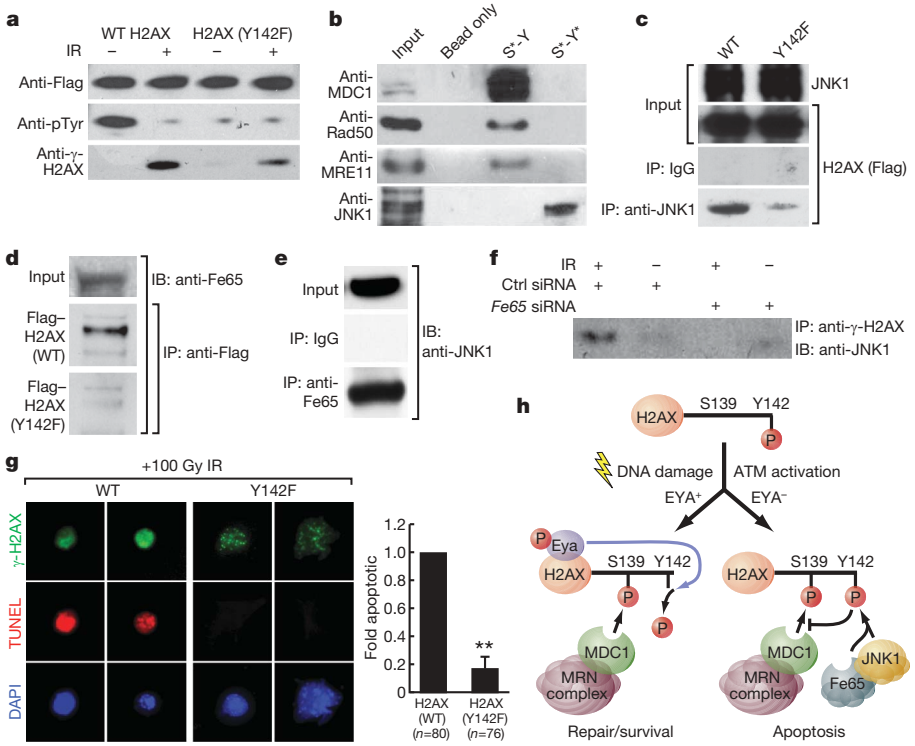


Figure 5 | H2AX Y142 phosphorylation discriminates between apoptotic and repair responses to DNA damage. **a**, S139 phosphorylation of H2AX Y142F is present but reduced in comparison to wild-type H2AX after 5 Gy ionizing radiation. **b**, Affinity purification performed on nuclear extract from irradiated 293T cells using synthetic peptides representing the C-terminal tail of H2AX bearing S139 phosphorylation with or without Y142 phosphorylation followed by western blot analysis. **c**, **d**, Co-immunoprecipitation confirms the interaction between wild-type H2AX and JNK1 (**c**) or Fe65 (**d**), but not H2AX Y142F, in 293T cells exposed to high-dose ionizing radiation (50 Gy). **e**, Endogenous Fe65 interacts with JNK1 in 293T cells treated with etoposide (30 μM). Panels **d** and **e** show individual bands from a single western blot exposure. **f**, siRNA knockdown of *Fe65* in 293T cells blocks the damage-dependent (50 Gy) interaction of JNK1 and γ-H2AX by co-immunoprecipitation in cells transfected with *Fe65* siRNA or control siRNA 48 h before harvest. Results were confirmed with two separate siRNA sets for *Fe65*. **g**, *H2ax*$^{-/-}$ MEFs were transfected with wild-type or mutant H2AX (Y142F) expression constructs and exposed to high-dose ionizing radiation (100 Gy). Apoptotic response among transfectants was assessed by γ-H2AX staining and TUNEL. Bar graphs represent mean ± s.e.m. of fold apoptotic values for triplicate or greater cell counts of transfected (green) nuclei. The basal level of apoptosis for wild-type H2AX transfected cells under these conditions was 25.7% TUNEL positive/total transfected nuclei. Values were normalized to wild-type H2AX-transfected samples. Double asterisk, $P < 0.001$. Original magnification, ×40. **h**, Proposed model for Y142 phosphorylation status of H2AX in regulation of apoptotic versus repair response.

DNA damage checkpoint protein 1 (MDC1), which has been shown to bind directly to phosphorylated S139 of H2AX at the sites of double-strand breaks[24] based on tandem BRCT1 repeats within the C terminus of MDC1 (ref. 25). MDC1 functions in the recruitment of a set of ancillary repair factors including MRE11, RAD50, NBS1 (the MRN complex), 53BP1 and BRCA1 (refs 26, 27), although these factors are not wholly dependent on MDC1 and γ-H2AX for recruitment to breaks[28]. Because an intact H2AX COOH-terminal tyrosine has been found to be required for MDC1–H2AX interaction and productive DNA repair[24], it was of particular interest to determine whether persistent phosphorylation of Y142 in the absence of EYA could have a negative impact on MDC1 recruitment to the tail of γ-H2AX. We first generated peptides corresponding to the C-terminal tail of H2AX with phosphorylation of both S129 and Y142, or of S139 alone. Peptides lacking any phosphorylation marks or where tyrosine 142 was mutated to alanine failed to interact with MDC1, consistent with previously published reports (Supplementary Fig. 9)[24]. Affinity purification of nuclear extract from irradiated 293T cells with each peptide revealed that, in the absence of Y142 phosphorylation, a set of DNA repair factors including MDC1, MRE11 and Rad50 were bound to the S139 phosphorylated H2AX peptide (Fig. 5b). Intriguingly, when phosphorylated tyrosine 142 was present with phosphoserine 139, binding of these factors was greatly reduced; instead, the established pro-apoptotic factor JNK1 was now present (Fig. 5b). The stress-response kinase JNK1, activated by DNA damage and initiating a pro-apoptotic program, has been recently shown to translocate into the nucleus on activation where it phosphorylates substrates including H2AX S139, an event critical for DNA degradation mediated by caspase-activated DNase (CAD) in apoptotic cells[10]. In agreement with our peptide purification experiments, we were able to detect a robust interaction between transfected wild-type H2AX and endogenous JNK1 in 293T cells in response to high-dose radiation; this interaction was markedly reduced in the case of the H2AX Y142F mutant (Fig. 5c).

To confirm further the specificity of these phosphorylation-dependent interactions we performed peptide competition assays. The H2AX tail peptide phosphorylated on S139 alone was able to compete effectively for binding of MDC1 in a peptide pull-down assay, whereas the free peptide bearing both S139 and Y142 phosphorylation marks competed away interaction with JNK1 (Supplementary Fig. 10).

On the basis of our previous data that loss of EYA phosphatase results in increased tyrosine phosphorylation of H2AX, we predicted that depleting EYA in 293T cells would result in decreased binding of MDC1 to H2AX in response to DNA damage. We knocked down EYA3 using specific siRNA and subsequently tested for MDC1–H2AX interaction by co-immunoprecipitation. As predicted, loss of EYA3 resulted in complete loss of this interaction in comparison to untransfected cells treated with 10 Gy ionizing radiation (Supplementary Fig. 11).

It was of particular interest to identify proteins containing SH2 and PTB phosphotyrosine-binding domains that could bind directly to H2AX phosphotyrosine 142 under conditions of genotoxic stress. We tested a set of known nuclear proteins containing these domains for binding to tyrosine-phosphorylated H2AX (Supplementary Table 1, partial list) and found that, whereas most exhibited no interaction, the PTB-domain protein Fe65[29], a cofactor for several cell-surface receptors that has been shown to translocate to the nucleus during DNA damage response and suggested to exert a pro-apoptotic role[30,31], bound specifically to wild-type γ-H2AX under DNA damage conditions, but not to the γ-H2AX Y142F mutant (Fig. 5d). Notably, we found that Fe65 protein interacted with endogenous JNK1 by co-immunoprecipitation in 293T cells treated with the DNA-damage agent etoposide (Fig. 5e), consistent with the idea that Fe65 helps to mediate JNK1 recruitment to γ-H2AX. Co-immunoprecipitation experiments demonstrated that the second PTB domain on Fe65 may be crucial for the interaction between Fe65 and tyrosine phosphorylated H2AX (Supplementary Fig. 12a). Glutathione *S*-transferase (GST) pull-down assays using purified recombinant protein of Fe65 PTB domains 1 and 2 also revealed a direct interaction between PTB2 and the H2AX present in purified HeLa histones (Supplementary Fig. 12b). We postulated that Fe65 may function as an adaptor protein, binding directly to the phosphotyrosine residue on γ-H2AX via PTB2 and facilitating the recruitment of pro-apoptotic factors such as JNK1. To test this, we knocked down endogenous Fe65 in 293T cells using specific siRNAs (Supplementary Fig. 2) and assessed the interaction between H2AX and JNK1 in response to genotoxic stress by co-immunoprecipitation. Whereas control siRNA had no effect on the ability of H2AX to co-immunoprecipitate JNK1, knockdown of Fe65 strongly inhibited this interaction (Fig. 5f).

To confirm the function of tyrosine 142 phosphorylation in regulation of the apoptotic response, we transfected *H2ax*$^{-/-}$ mouse embryonic fibroblasts (MEFs)[32] with either wild-type or Y142F H2AX expression constructs. When these cells were subjected to high-dose ionizing radiation, cells expressing H2AX Y142F displayed a reduced apoptotic response in comparison to cells expressing wild-type H2AX (~6-fold decrease) (Fig. 5g). These data suggested to us that lack of H2AX Y142 phosphorylation promotes a damage repair response instead of an apoptotic response to DNA damage, in part by promoting successful recruitment of MDC1 and associated repair factors. The presence of Y142 phosphorylation in wild-type H2AX transfected MEFs is proposed to lead to the recruitment of pro-apoptotic factors such as JNK1 to H2AX, while inhibiting the recruitment of the damage repair complex, directly promoting apoptotic response to genotoxic stress.

Conclusions

Cells are confronted with DNA damage resulting from a variety of stimuli under normal physiological conditions and at each instance the cell must make fundamental decisions concerning the ratio of DNA repair and apoptotic response. Our data suggest that γ-H2AX is involved in the adjudication of the balance between these two outcomes, with a single post-translational modification, phosphorylation of tyrosine 142, being capable of influencing the recruitment to γ-H2AX of functional apoptotic or repair complexes. In the presence of Y142 phosphorylation, binding of repair factors to phosphorylated serine 139, which is mediated by MDC1, is inhibited (Fig. 5h), whereas recruitment of pro-apoptotic factors, including JNK1, is promoted.

EYA binds to SIX-class homeodomain transcription factors. Although early *in vitro* studies suggested that phosphatase activity was important for EYA-mediated transcriptional activation of certain SIX-dependent reporter genes[6], recent studies in *Drosophila* suggest that most Six/Eya transcriptional targets do not require phosphatase enzymatic activity for activation *in vivo*[33]. Phosphatase activity of EYA may have a novel function in mammalian organogenesis, acting to block an improper apoptotic response to physiological levels of genotoxic stress by dephosphorylating H2AX on tyrosine.

Coincident with our studies, recently published work reported phosphorylation of H2AX on tyrosine 142 under basal conditions which decreases in response to DNA damage in MEFs[34]. The relevant kinase was demonstrated to be WSTF (Williams–Beuren syndrome transcription factor), which physically interacts with H2AX specifically in undamaged cells. The authors demonstrated that siRNA knockdown of WSTF results in loss of H2AX Y142 phosphorylation, which alters the kinetics of S139 phosphorylation in response to DNA damage. Thus, it seems that H2AX tyrosine phosphorylation is deposited by WSTF under basal conditions and, at least in the embryonic kidney cell model system, is removed by EYA in response to DNA damage.

The present study indicates that the phosphorylation of tyrosine 142 of H2AX prevents recruitment of repair complexes to phosphoserine 139 of γ-H2AX, although it is likely that there are many additional aspects that underlie the full molecular logic for the dual

phosphorylation-mediated events. We hypothesize that the presence of both phosphorylated residues results in direct binding of the PTB domain factor Fe65, which, at least in part, mediates the effective recruitment of other pro-apoptotic factors, including JNK1.

METHODS SUMMARY

Eya1 knockout mice were originally generated by the laboratory of R. Maas. 293T and *H2ax*$^{-/-}$ MEF cells were maintained in DMEM (Gibco) supplemented with 10% fetal calf serum (FCS; Gemini). Plasmids and siRNAs were transfected with Lipofectamine 2000 (Invitrogen) as directed. Specific antibodies for immunoprecipitation and immunostaining were obtained from Upstate (anti-γ-H2AX), Zymed (anti-phosphotyrosine), Cell Signaling Technology (anti-H2AX, anti-γ-H2AX), Abcam (anti-KSP-cadherin 16, anti-MDC1), Sigma (anti-Flag), and Santa Cruz Biotechnology (anti-RAD50, MRE11, JNK1). Purified peptides were obtained from Sigma Genosys, Abgent, and Anaspec.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 December 2008; accepted 4 February 2009.
Published online 22 February 2009.

1. Kumar, J. P. Signalling pathways in *Drosophila* and vertebrate retinal development. *Nature Rev. Genet.* **2,** 846–857 (2001).
2. Silver, S. J. & Rebay, I. Signaling circuitries in development: insights from the retinal determination gene network. *Development* **132,** 3–13 (2005).
3. Pignoni, F. *et al.* The eye-specification proteins So and Eya form a complex and regulate multiple steps in *Drosophila* eye development. *Cell* **91,** 881–891 (1997).
4. Bonini, N. M., Leiserson, W. M. & Benzer, S. The eyes absent gene: genetic control of cell survival and differentiation in the developing *Drosophila* eye. *Cell* **72,** 379–395 (1993).
5. Xu, P. X. *et al. Eya1*-deficient mice lack ears and kidneys and show abnormal apoptosis of organ primordia. *Nature Genet.* **23,** 113–117 (1999).
6. Li, X. *et al.* Eya protein phosphatase activity regulates Six1–Dach–Eya transcriptional effects in mammalian organogenesis. *Nature* **426,** 247–254 (2003).
7. Rayapureddi, J. P. *et al.* Eyes absent represents a class of protein tyrosine phosphatases. *Nature* **426,** 295–298 (2003).
8. Tootle, T. L. *et al.* The transcription factor Eyes absent is a protein tyrosine phosphatase. *Nature* **426,** 299–302 (2003).
9. Rayapureddi, J. P. *et al.* Characterization of a plant, tyrosine-specific phosphatase of the aspartyl class. *Biochemistry* **44,** 751–758 (2005).
10. Lu, C. *et al.* Cell apoptosis: requirement of H2AX in DNA ladder formation, but not for the activation of caspase-3. *Mol. Cell* **23,** 121–132 (2006).
11. Bassing, C. H. & Alt, F. W. The cellular response to general and programmed DNA double strand breaks. *DNA Repair* **3,** 781–796 (2004).
12. Bassing, C. H. *et al.* Increased ionizing radiation sensitivity and genomic instability in the absence of histone H2AX. *Proc. Natl Acad. Sci. USA* **99,** 8173–8178 (2002).
13. Karagiannis, T. C. & El-Osta, A. Chromatin modifications and DNA double-strand breaks: the current state of play. *Leukemia* **21,** 195–200 (2007).
14. van Attikum, H. & Gasser, S. M. The histone code at DNA breaks: a guide to repair? *Nature Rev. Mol. Cell Biol.* **6,** 757–765 (2005).
15. Lee, Y. M. *et al.* Determination of hypoxic region by hypoxia marker in developing mouse embryos *in vivo*: a possible signal for vessel development. *Dev. Dyn.* **220,** 175–186 (2001).
16. Haase, V. H. Hypoxia-inducible factors in the kidney. *Am. J. Physiol. Renal Physiol.* **291,** F271–F281 (2006).
17. Fernandez-Capetillo, O. *et al.* H2AX: the histone guardian of the genome. *DNA Repair* **3,** 959–967 (2004).
18. Rogakou, E. P. *et al.* Megabase chromatin domains involved in DNA double-strand breaks *in vivo*. *J. Cell Biol.* **146,** 905–916 (1999).
19. Berkovich, E., Monnat, R. J. Jr & Kastan, M. B. Roles of ATM and NBS1 in chromatin structure modulation and DNA double-strand break repair. *Nature Cell Biol.* **9,** 683–690 (2007).
20. Berkovich, E., Monnat, R. J. Jr & Kastan, M. B. Assessment of protein dynamics and DNA repair following generation of DNA double-strand breaks at defined genomic sites. *Nature Protocols* **3,** 915–922 (2008).
21. Matsuoka, S. *et al.* ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316,** 1160–1166 (2007).
22. Stokes, M. P. *et al.* Profiling of UV-induced ATM/ATR signaling pathways. *Proc. Natl Acad. Sci. USA* **104,** 19855–19860 (2007).
23. Lavin, M. F. & Kozlov, S. ATM activation and DNA damage response. *Cell Cycle* **6,** 931–942 (2007).
24. Stucki, M. *et al.* MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks. *Cell* **123,** 1213–1226 (2005).
25. Lee, M. S. *et al.* Structure of the BRCT repeat domain of MDC1 and its specificity for the free COOH-terminal end of the γ-H2AX histone tail. *J. Biol. Chem.* **280,** 32053–32056 (2005).
26. Kim, J. E., Minter-Dykhouse, K. & Chen, J. Signaling networks controlled by the MRN complex and MDC1 during early DNA damage responses. *Mol. Carcinog.* **45,** 403–408 (2006).
27. Wu, X. *et al.* ATM phosphorylation of Nijmegen breakage syndrome protein is required in a DNA damage response. *Nature* **405,** 477–482 (2000).
28. Celeste, A. *et al.* Histone H2AX phosphorylation is dispensable for the initial recognition of DNA breaks. *Nature Cell Biol.* **5,** 675–679 (2003).
29. Duilio, A. *et al.* A rat brain mRNA encoding a transcriptional activator homologous to the DNA binding domain of retroviral integrases. *Nucleic Acids Res.* **19,** 5269–5274 (1991).
30. Minopoli, G. *et al.* Essential roles for Fe65, Alzheimer amyloid precursor-binding protein, in the cellular response to DNA damage. *J. Biol. Chem.* **282,** 831–835 (2007).
31. Nakaya, T., Kawai, T. & Suzuki, T. Regulation of FE65 nuclear translocation and function by amyloid β-protein precursor in osmotically stressed cells. *J. Biol. Chem.* **283,** 19119–19131 (2008).
32. Celeste, A. *et al.* Genomic instability in mice lacking histone H2AX. *Science* **296,** 922–927 (2002).
33. Jemc, J. & Rebay, I. Identification of transcriptional targets of the dual-function transcription factor/phosphatase eyes absent. *Dev. Biol.* **310,** 416–429 (2007).
34. Xiao, A. *et al.* WSTF regulates the H2A.X DNA damage response via a novel tyrosine kinase activity. *Nature* **457,** 57–62 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Kastan for providing reagents/technical assistance for the I-PpoI system. We thank V. Lunyak, J. Dixon and R. Koladner for review and discussions. We thank the laboratory of R. S. Johnson for use of equipment and advice on hypoxia incubations, as well as H. Taylor for animal care assistance and C. Nelson for cell culture assistance. We thank A. Nussenzweig, Y. Xu and H. Song for *H2ax*$^{-/-}$ MEFs. We thank J. Hightower and M. Fisher for assistance with figure and manuscript preparation. We additionally thank X. Li and W. Liu. M.G.R. is an HHMI Investigator. This work was supported by grants from NIH and NCI to M.G.R. and C.K.G. This work also was supported by the Sogang University Research Grant of 2008 to B.G.J and PCF and USAMRAA grants to M.G.R.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.G.R. (mrosenfeld@ucsd.edu).

METHODS

Antibodies, reagents and cells. The following commercially available antibodies were used: anti-H2AX (Cell Signaling Technology and Abcam), anti-γ-H2AX (Cell Signaling Technology and Upstate), anti-phosphotyrosine (Zymed and Upstate), anti-KSP-cadherin 16 (Abcam), anti-HA (Berkeley Antibody Company), anti-Flag (Sigma), anti-MDC1 (Abcam and Bethyl laboratories), anti-RAD50, MRE11, JNK1 (Abcam and Santa Cruz Biotechnology). Antibodies to EYA3 were generated by immunizing guinea-pigs with GST-purified peptides representing the amino terminus of human EYA3 (amino acids 1–239). The following commercially available reagents were used: caffeine (Calbiochem). *EYA1* and *EYA3* siRNAs were purchased from Qiagen. *H2ax*$^{-/-}$MEFs were provided by A. Nussenzweig, Y. Xu and H. Song. Standard molecular cloning and tissue culture were performed as described[35].

Animal care and immunohistochemistry. *Eya1* knockout mice were originally generated by the laboratory of R. Mass. Mouse embryos from E10.5 to E11.5 were fixed in 2% paraformaldehyde, penetrated with 24% sucrose in PBS, and embedded in OCT compound for cryo-sectioning. Serial 14-μm sections were blocked in 10% normal goat serum/PBS/0.1% Triton X-100 and immunostained using antibodies to γ-H2AX or KSP-cadherin 16. Immunostaining was visualized using secondary antibodies conjugated to Alexa-Fluor-595 (Invitrogen) and sections were mounted using Vectashield mounting media plus DAPI (Vector Laboratories). Parallel sections were stained with haematoxylin and eosin as described[6].

TUNEL staining. TUNEL assay was performed using ApopTag *In situ* apoptosis detection kit (Chemicon). Tissue sections were post-fixed in ethanol:acetic acid 2:1 at −20 °C for 5 min and incubated with TdT enzyme at 37 °C for 1 h. DIG incorporation was visualized using anti-digoxigenin-rhodamine secondary (Roche) and stained sections were mounted using Vectashield mounting media plus DAPI (Vector Laboratories).

Cell treatment and transfection/RNA interference. For hypoxia experiments, 293T cells were transferred to an 8% CO_2, 2% O_2 incubator and maintained for approximately 20 h. Cells were immediately fixed or lysed on removal from the hypoxia incubator. Gamma-irradiation of cultured cells was performed at the UCSD Medical Teaching Facility according to established protocols. The cells were gamma-irradiated approximately 36–48 h after transfection. Cells were transfected using Lipofectamine 2000 (Invitrogen). siRNA target sequences were as follows: *EYA1*, CAGGAAATAATTCACTCACAA; *EYA3*, CCGGAAAGTGA GAGAAATCTA; *Fe65*, CTGTATTGATATCACTAATAA (Qiagen), CUACGUA GCUCGUGAUAAG, GGGUAGAUGUGAUUAAUGG, GAUCAAGUGUUUC GCCGUG, CGUCAGCUCUCUUACCACA (Dharmacon).

Immunoprecipitation/western blot analysis. For immunoprecipitation and western blotting, cells were rinsed in PBS, harvested and lysed in lysis buffer containing 10% glycerol, 0.5 mM EDTA, 25 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM Na_2VO_3, 10 mM β-glycerophosphate, 0.1% NP-40 and 1 mM dithiothreitol in the presence of protease inhibitors (Roche) and 1 mM PMSF. The extracts were incubated with the specific antibody overnight at 4 °C, followed by incubation with protein A/G agarose beads (Santa Cruz Biotech), washed extensively, and separated by electrophoresis. Proteins were transferred onto nitrocellulose membranes (Bio-Rad) and western blotting was performed following standard protocols.

Immunocytochemistry. Cells were fixed for 15 min with 2% paraformaldehyde in PBS and permeabilized with 0.05% Triton X-100 in PBS for 30 min. After blocking with PGBA solution (0.1% BSA, 0.1% gelatine, 0.1% FBS), cells were incubated with specific antibodies for 2 h at room temperature. Antigen was detected with secondary antibodies conjugated to Alexa-Fluor-595 or Alexa-Fluor-488 (Invitrogen). Cells were coverslipped using Vectasheild mounting media plus DAPI (Vector Laboratories).

***In vitro* phosphatase assay.** The HA-tagged EYA phosphatase was immunoprecipitated from gamma-irradiated 293T cells using anti-HA affinity resin (Roche). After extensive washing, EYA phosphatase was eluted with HA peptide. The reaction mixture containing purified EYA protein in 100 μl phosphatase buffer (50 mM Tris-HCl, pH 7.0, 5 mM $MgCl_2$, 10% glycerol, 3 mg ml^{-1} BSA) and bovine histone (Sigma) was incubated for 60–90 min at 30 °C. H2AX was immunoprecipitated with anti-H2AX antibody and western blotting was performed. GST fusion proteins of EYA3 240–573 and EYA3 D246A 240–573 were expressed in BL21 bacterial cells and purified with glutathione-agarose beads (Sigma). Wild-type and mutant GST proteins were incubated with 2 mg purified peptides of the H2AX tail bearing phosphorylation at either S139 or Y142 (Abgent) in phosphatase buffer for 1 h and free phosphatase was detected using Malachite Green (BIOMOL).

Peptide affinity chromatography. Biotinylated synthetic peptides (hH2AX amino acids 128–142) were purchased from Sigma Genosys, Anaspec and Abgent. For peptide affinity chromatography, biotinylated phosphopeptides and unphosphorylated peptides were coupled to streptavidin-coated Dynabeads M-280 (Invitrogen) for 2 h at room temperature. Beads were incubated with nuclear extract from 200-Gy-irradiated 293T cells and washed extensively with Tris buffered saline (pH 7.5) containing 0.5% Tween 20. The bound proteins were separated by SDS–PAGE using 4–12% Bis-Tris NuPAGE gel (Invitrogen), followed by western blot analysis.

35. Sambrook, J. & Russell, D. W. *Molecular Cloning: a Laboratory Manual* 3rd edn (Cold Spring Harbor Laboratory Press, 2001).

ARTICLES

Structure of the connexin 26 gap junction channel at 3.5 Å resolution

Shoji Maeda[1], So Nakagawa[1], Michihiro Suga[1], Eiki Yamashita[1], Atsunori Oshima[2], Yoshinori Fujiyoshi[2] & Tomitake Tsukihara[1,3]

Gap junctions consist of arrays of intercellular channels between adjacent cells that permit the exchange of ions and small molecules. Here we report the crystal structure of the gap junction channel formed by human connexin 26 (Cx26, also known as GJB2) at 3.5 Å resolution, and discuss structural determinants of solute transport through the channel. The density map showed the two membrane-spanning hemichannels and the arrangement of the four transmembrane helices of the six protomers forming each hemichannel. The hemichannels feature a positively charged cytoplasmic entrance, a funnel, a negatively charged transmembrane pathway, and an extracellular cavity. The pore is narrowed at the funnel, which is formed by the six amino-terminal helices lining the wall of the channel, which thus determines the molecular size restriction at the channel entrance. The structure of the Cx26 gap junction channel also has implications for the gating of the channel by the transjunctional voltage.

Intercellular signalling is one of the most essential properties of multicellular organisms. Gap junctions are specialized membrane regions containing hundreds of intercellular communication channels that allow the passage of molecules such as ions, metabolites, nucleotides and small peptides[1]. A gap junction channel is formed by end-to-end docking of two hemichannels, also referred to as connexons, each composed of six connexin subunits[2]. Connexin is predicted to have four transmembrane helices and two extracellular loops, which are thought to contain a β-strand structure and are an essential structural basis for the docking of two connexons[3]. Gap junctions have crucial roles in many biological processes including development, differentiation, cell synchronization, neuronal activity and immune responses[4,5]. Mutations in connexins thus cause several human diseases, including neurodegenerative diseases, skin diseases, deafness and developmental abnormalities[5,6].

To date, more than 20 different connexins have been identified in the human genome, which have been categorized into α, β and γ isoforms on the basis of their sequence homology. The connexin composition of gap junction channels defines their unique properties, such as their selectivity for small molecules, voltage-dependent gating, and response to Ca^{2+}, pH and phosphorylation[5,7].

Early electron microscopic analyses of gap junctions suggested that channel gating involves a rotation of all six subunits[8,9], and analysis of two-dimensional crystals formed by carboxy-terminally truncated connexin 43 (Cx43, also known as GJA1) resulted in a model for the arrangement of the transmembrane helices and the fold of the connexin protomer[10,11]. Recently, the electron crystallographic analysis of the connexin 26 Met34Ala mutant (Cx26(M34A)) revealed large densities in the pore at the level of the two membranes, which were interpreted as plugs blocking the channel[12]. The structure of Cx26(M34A) was thus assumed to show the channel in a closed state. The structure also suggested that physical blockage by a plug is an essential part of a gating mechanism and is consistent with the physiological studies showing that each connexon can regulate its activity autonomously[13–15]. Electrophysiological studies have demonstrated that gap junctions have several gating mechanisms. At least two regulation mechanisms respond to the transjunctional voltage (V_j), V_j gating (fast) and loop gating (slow)[16]. Gap junctions can also be gated by the membrane voltage (V_m), termed V_m gating, and by chemical factors such as phosphorylation, pH and Ca^{2+}, known as chemical gating[17].

Here we present an atomic structure of the human Cx26 gap junction channel. We find that the four transmembrane helices of a protomer are arranged differently from the previously proposed pseudoatomic model[11], and that several residues associated with non-syndromic hereditary deafness or skin diseases are involved in intra- or intermolecular interactions. We describe in detail the interactions between the two extracellular regions of adjoining connexons. The N-terminal regions of the six subunits line the pore entrance and form a funnel, which restricts the diameter at the entrance of the pore to 14 Å. In conjunction with previous electron microscopy work[12], this finding suggests that conformational changes in the Cx26 N termini play an important part in channel gating, specifically in V_j gating.

Structure determination of the gap junction channel

Structure determination at 3.5 Å is briefly described in the Methods. The whole structure of each protomer—except for residues 110–124 and 218–226 that correspond to most of the cytoplasmic loop and the carboxy-terminal segment, respectively—was successfully modelled in electron density maps. The amino acid assignment was confirmed by methionine sites and disulphide bonds sites (Supplementary Fig. 1). Of the 226 residues of Cx26, the atomic parameters of residues 2–109 and 125–217 converged well during refinement.

The overall structure of the Cx26 gap junction channel, which is formed by two connexons related to each other by a crystallographic two-fold symmetry axis, is similar in shape and size to that of the C-terminal truncated Cx43 gap junction channel visualized by electron crystallography[10] (Fig. 1a). It is a tsuzumi shape, a traditional Japanese drum. The protomers in each hexameric connexon are related by a six-fold non-crystallographic symmetry (NCS) axis perpendicular to the membrane plane (Fig. 1b). The height of the modelled structure of the gap junction channel without disordered cytoplasmic loop and

[1]Institute for Protein Research, Osaka University, OLABB, 6-2-3, Furuedai, Suita, Osaka 565-0874, Japan. [2]Department of Biophysics, Graduate School of Science, Kyoto University, Oiwake, Kitashirakawa, Sakyo-ku, Kyoto 606-8502, Japan. [3]Picobiology Institute, Graduate School of Life Science, University of Hyogo, Kamigohori, Akoh, Hyogo 678-1297, Japan.

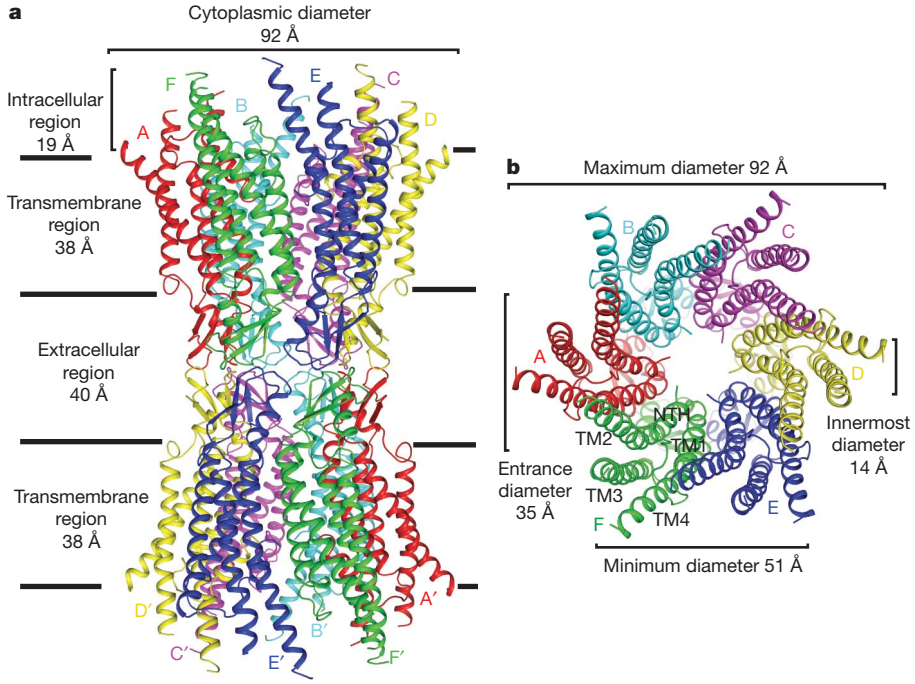


Figure 1 | Overall structure of the Cx26 gap junction channel in ribbon representation. The corresponding protomers in the two hemichannels, which are related by a two-fold axis, are shown in the same colour. **a**, Side view of the Cx26 gap junction channel. **b**, Top view of the Cx26 gap junction channel showing the arrangement of the transmembrane helices TM1 to TM4. The pore has an inner diameter of 35 Å at the cytoplasmic entrance, and the smallest diameter of the pore is 14 Å.

C-terminal segment is approximately 155 Å. The transmembrane region and membrane surfaces were deduced from the distribution of hydrophobic and aromatic amino acid residues along the non-crystallographic six-fold axis (Fig. 1a and Supplementary Fig. 2). The transmembrane region of the channel is 38 Å thick. TM2 extends about 19 Å from the membrane surface into the cytoplasm. The extracellular region of the connexon extends 23 Å from the membrane surface and interdigitates to the opposite connexon by 6 Å, resulting in the intercellular 'gap' of 40 Å. The extracellular lobes are not protruding so much, as indicated by the structural analyses of split gap junction channels with atomic force microscopy and electron microscopy[18,19]. The relatively flat lobes could be attributed to the conformational change of the extracellular region induced by the docking of two connexons. The diameter of the connexon is biggest at the cytoplasmic side of the membrane, ~92 Å, and smallest at the extracellular side, ~51 Å. Viewed from the top, the channel looks like a 'hexagonal nut' with a pore in the centre (Fig. 1b). The diameter of the pore is about 40 Å at the cytoplasmic side of the channel, narrowing to 14 Å near the extracellular membrane surface and then widening to 25 Å in the extracellular space. No obvious obstructions are detectable throughout the solute pathway, although this does not exclude the possibility that the cytoplasmic domains not resolved in our map may be able to form a gate. Because the 3.5 Å X-ray structure does not show any obstructions along the pore, our structure of wild-type Cx26 seems to be in an open conformation, which is consistent with the crystallization conditions used (neutral pH without aminosulphonate buffer or any divalent ions).

Structure of the Cx26 protomer

The protomer has four transmembrane segments (TM1–4), two extracellular loops (E1 and E2), a cytoplasmic loop, an N-terminal helix (NTH), and a C-terminal segment (Fig. 2 and Supplementary Fig. 3). Cx26 forms a typical four-helix bundle in which any pair of adjacent helices is antiparallel. TM1 and TM2 face the interior, whereas TM3 and TM4 face the hydrophobic membrane environment. There has been controversy about the identity of the major pore-lining helix, on the basis of accessibility studies of substituted cysteines and sequence analysis. One set of data favours TM3 as the major pore helix[11,20] and the other favours TM1 (refs 21, 22). The helical arrangement of our structure is consistent with the latter model. The major pore-lining helix TM1 is inclined, so that the pore diameter narrows from the cytoplasmic to the extracellular side of the membrane, and ends in a short 3_{10} helix (Fig. 2 and Supplementary Fig. 4). TM2 is kinked at Pro 87, the midpoint of the helix, and TM2 and TM3 protrude into the cytoplasm. The Pro87Leu mutation has been shown to cause an aberrant gating[23]. Furthermore, mutations of three residues to proline (Leu79Pro, Ser85Pro and Leu90Pro) in TM2 link to deafness[24]. These mutations probably evoke a structural change in TM2, which would affect the cytoplasmic domains including the NTH. TM4 inclines from the molecular axis by about 30°, generating a larger diameter of the connexon on the intracellular side.

The extracellular loop E1 contains a 3_{10} helix at the beginning and a short α-helix in its C-terminal half (Fig. 2 and Supplementary Fig. 3). E2, together with E1, contains a short antiparallel β-sheet and stretches over E1, forming the outside wall of the connexon. Six conserved cysteine residues, three in each loop, form intramolecular disulphide bonds between E1 and E2 (ref. 3) (Figs 2, 3a and Supplementary Fig. 1). The N-terminal half of E2 seems rather flexible and its amino-acid sequence varies greatly among connexins (Supplementary Fig. 5). The C-terminal half of E2 begins with a 3_{10} turn and is followed by a conserved Pro-Cys-Pro motif that reverses its direction back to TM4.

Most of the prominent intra-protomer interactions are in the extracellular part of the transmembrane region (Fig. 3a and Supplementary Fig. 6). Arg 32 (TM1) interacts with Gln 80 (TM2), Glu 147 (TM3), and Ser 199 (TM4). Two hydrophobic cores around Trp 44 (E1) and Trp 77 (TM2) stabilize the protomer structure. Ala 39 (TM1), Ala 40 (TM1), Val 43 (E1) and Ile 74 (TM2) contribute to the first hydrophobic core around Trp 44, and Phe 154 (TM3) and Met 195 (TM4) form the second core with Trp 77 (Supplementary Fig. 6). In the intracellular part of the transmembrane region, Arg 143 (TM3) forms hydrogen bonds with Asn 206 (TM3) and

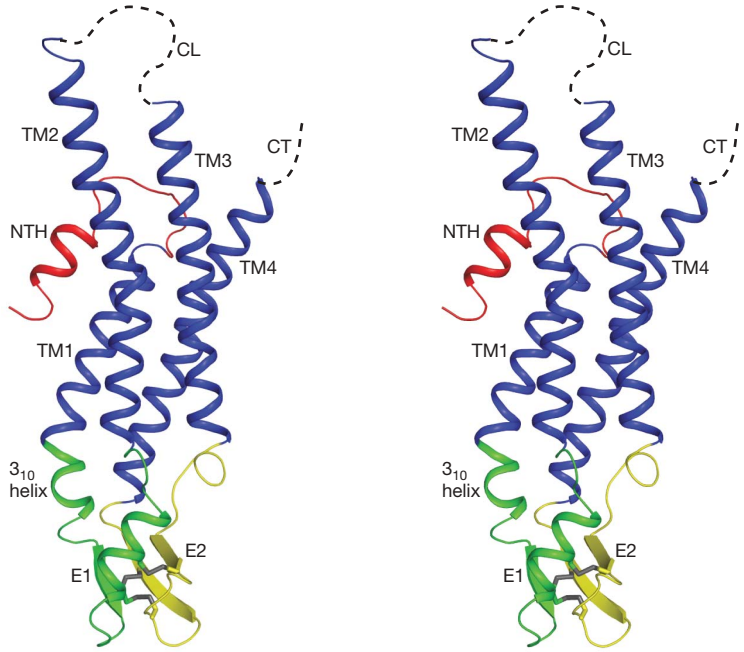


Figure 2 | Stereo view of the Cx26 protomer in ribbon representation. Colour code: red, NTH; blue, TM1–TM4; green, E1; yellow, E2; grey, disulphide bonds; dashed lines, cytoplasmic loop (CL) and C terminus (CT), which were not visible in the map. E1 and E2 are the loops connecting TM1 and TM2, and TM3 and TM4, respectively.

Ser 139 (TM3) (Supplementary Fig. 6). The four-helix bundle is further stabilized by dipole–dipole interactions of the antiparallel helices[25].

Structural organization of the hexameric connexon

The inter-protomer interactions in the hexameric connexon are mostly located in the extracellular half of transmembrane helices TM2 and TM4 and in the extracellular loops. Glu 47 (E1), Gln 48 (E1), Asn 62 (E1), Asp 66 (E1), Tyr 65 (E1), Arg 75 (TM2) and the main-chain amide of Ser 72 (E1) from one protomer, and Asp 46 (E1), Asp 50 (E1), Arg 184 (E2) Thr 186 (TM4) and Glu 187 (TM4) from the adjacent protomer form the core of the inter-protomer interactions (Supplementary Fig. 7). Although TM3 is evolutionarily more variable than the other three helices, every third or fourth residue in TM3 is aromatic, generating an aromatic face that is conserved among connexins. Each helix in a protomer contributes to an aromatic cluster in the groove between two adjacent protomers (Supplementary Fig. 7). Most of the residues involved in intra- and inter-protomer interactions are conserved

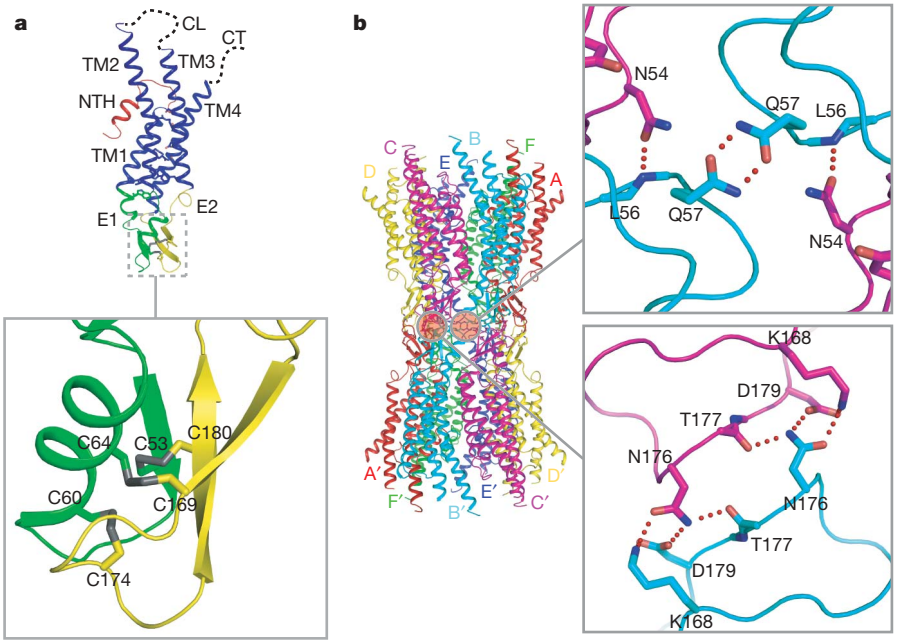


Figure 3 | Molecular architecture of the Cx26 gap junction channel. The Cα trace is shown in ribbon or line representation and the side chains in the close-up views in the boxes are shown as sticks. Hydrogen bonds or salt bridges are shown as dotted lines. **a**, Disulphide bonds between two extracellular loops in the Cx26 protomer. **b**, Intercellular interactions. The protomers forming the gap junction channel are labelled A to F and A′ to F′ each in the same colour as in Fig. 2. The right top and bottom boxes show intercellular interactions in E1 and E2, respectively.

within the connexin family (Supplementary Fig. 5), and mutations of these residues are associated with deafness and skin diseases[24]. The mutations probably interfere with the proper folding and/or oligomerization of connexins, thus resulting in defective channels.

Architectures of the intercellular junction and channel

Our structure revealed the interactions between the two adjoining connexons of the gap junction channel, which involve both E1 and E2 (Fig. 3b). In E1, Asn 54 forms hydrogen bonds with the main-chain amide of Leu 56 in the opposite protomer, and Gln 57 forms symmetric hydrogen bonds with the same residue of the diagonally opposite protomer. These residues are highly conserved among connexins (Supplementary Fig. 5). In E2, Lys 168, Asp 179 and the main-chain carbonyl groups of Thr 177 and Asn 176 form hydrogen bonds and salt bridges with the opposite protomer. Together with interactions between the protomers in the two hemichannels, these interactions create a tight double-layered wall bridging the extracellular gap, which connects the two adjoining hemichannels and separates the channel interior from the extracellular environment.

The permeation pathway of a gap junction channel consists of an intracellular channel entrance, a pore funnel and an extracellular cavity. The intracellular channel entrance has a diameter of 40 Å and is formed by the intracellular parts of TM2 and TM3. Eleven positively charged residues, nine in TM2 and two in TM3, generate a positively charged environment at the channel entrance (Fig. 4a). The positive atmosphere around the intracellular channel entrance would be favourable for concentrating and increasing absolute permeability of negatively charged molecules[26].

The funnel surface is lined by N-terminal residues Asp 2, Trp 3, Thr 5, Leu 6 and Ile 9 (Fig. 4b). Most α-connexins have a conserved Phe residue at the position of Thr 5 in Cx26, except for Cx43, which has an Ala residue (Supplementary Fig. 8). Because the funnel forms a constriction site at the cytoplasmic entrance of the pore, the size and electrical character of the side chains in this region should have a strong effect on both the molecular cutoff size and the charge selectivity of the channel. In line with this notion, it has previously been reported that the charges in the N-terminal region have a crucial involvement in determining the charge selectivity of the channel[27]. Cx43 channels are known to have the widest functional pore, followed by β-connexins and then other α-connexins[26,28], which could be reasonably derived from the size of the side chain at the position 5.

Twelve copies of the N-terminal half of E1 form the inner wall of the extracellular cavity of the pore, which has dimensions of $25 \times 25 \times 30$ Å^3 (Fig. 4a, b). This finding is in agreement with a functional study that demonstrated that E1 lines the pore in the extracellular gap region[29]. The pore-lining residues at the TM1/E1 boundary are Lys 41, Glu 42 and Gly 45. Lys 41 creates a narrowed part of the pore with the diameter of about 17 Å and is unique to Cx26 (Supplementary Fig. 8), generating a more positively charged environment between the funnel and the following negatively charged part of the solute pathway. The TM1/E1 boundary has been suggested to be involved in voltage sensing, together with the N terminus[14]. Although there is no direct interaction between Lys 41 and the N terminus of Cx26 (the distance between Lys 41 and the bottom of the funnel is approximately 8 Å), it is conceivable that Lys 41 and the Cx26-specific N terminus act together in sensing the voltage field. Asp 46 and Asp 50, highly conserved residues in the connexin family (Supplementary Fig. 8), face the pore interior and create a 9-Å long, negatively charged path with a diameter of 20 Å, approximately at the height of the extracellular membrane surface (Fig. 4a). Along with the pore funnel, these two regions probably contribute to the size restriction and possibly to the charge selectivity, considering the pore diameter and the charge character.

Pore funnel and the voltage-dependent gating mechanism

The short NTHs of the six protomers form the funnel (Fig. 5), and their very high crystallographic temperature factors indicate that these are the most mobile domains in the structure (Supplementary Fig. 9a, b). This finding agrees with an NMR solution structure of an N-terminal peptide of Cx26, which showed that the loop connecting the NTH to TM1 is very flexible[30]. Asp 2 forms hydrogen bonds with the main-chain amide of Thr 5 from the neighbouring protomer. The Asp 2 and Thr 5 residues on neighbouring NTHs at the bottom of the funnel form a circular girdle, as previously seen in the nicotinic acetylcholine receptor[31], which stabilizes the funnel structure (Fig. 5). Trp 3 forms

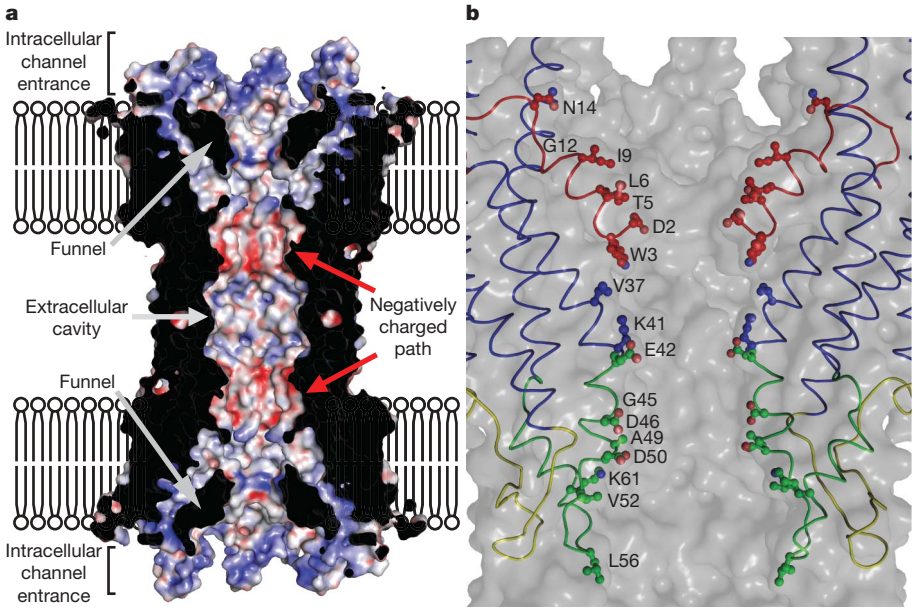


Figure 4 | Pore structure of the Cx26 gap junction channel. **a**, Vertical cross-section through the gap junction channel, showing the surface potential inside the channel. The channel features a wide cytoplasmic opening, which is restricted by the funnel structure, a negatively charged path and an extracellular cavity at the middle. Electrostatic surface potential of the Cx26 gap junction channel was calculated by the program APBS[43] as implemented in PyMOL under dielectric constants of 2.0 and 80.0 for protein and solvent regions, respectively. The displayed potentials range from −40 (red) to 40 (blue) kTe^{-1}. **b**, Pore-lining residues in a Cx26 gap junction channel. Side view of Cx26 gap junction channel pore; the main chain is depicted as a thin ribbon and side chains facing the pore as balls and sticks. For fine viewing, two subunits in the foreground are omitted in the surface representation and two further subunits in the background are omitted in the model depiction. The colouring is the same as in Fig. 3b.

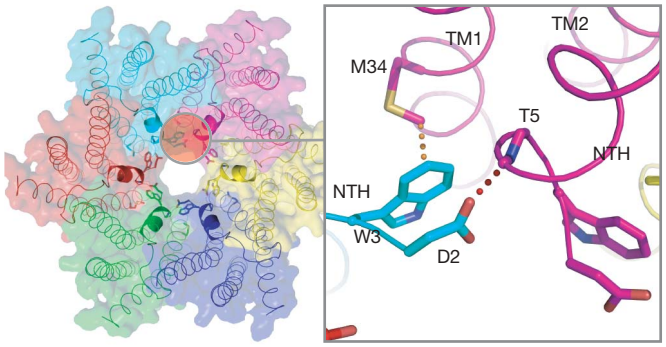


Figure 5 | Structure of the pore funnel. The six NTHs form a funnel structure, which is stabilized by a circular network of hydrogen bonds between Asp 2 and the main chain of Thr 5. The Cx26 protomers are shown in line and the NTHs in ribbon representation superposed on a surface representation. The close-up view shows the interaction between the indole ring of Trp 3 and the methyl group of Met 34 (TM1) in the adjacent protomer (hydrophobic interaction: orange broken line; hydrogen bond: red broken line).

hydrophobic interactions with Met 34 (TM1) of the neighbouring protomer, which draws the NTH to the inner wall of the channel. This interaction maintains the funnel in the open state, with an inner diameter of 14 Å. One of the most frequent deafness mutations is Met34Thr, which decreases electrical current, but forms structures indistinguishable from wild-type gap junctions[32,33]. This mutation would indeed disrupt the interaction of the NTH with Trp 3, which would cause the funnel to detach from the inner wall of the pore, resulting in a narrower funnel. This concept is supported by recent electron microscopy studies that showed a prominent density in the centre of the pore in Cx26(Met34Ala)[12], which was decreased in the N-terminal deletion mutant Cx26(Met34Ala-del2–7)[34].

Cx26 channels are known to be closed by an inside positive potential[14]. This is opposite from the gating property of Cx32, which has Asn at position 2 and closes after an inside negative potential[14]. A cytoplasmic movement of the N-terminal portion, where the voltage sensor is believed to reside, has been suggested to initiate voltage-dependent gating[14,35,36]. The recent electron microscopy structure of the Met34Ala mutant of Cx26 shows a plug that blocks the pore[12], which may be due to the smaller side chain at position 34 causing the channel to adopt a closed conformation. Although this electron microscopy structure may not exactly represent a physiological closed state, it is conceivable that an inside positive V_j would cause an inward movement of Asp 2, thus preventing the interactions between Asp 2–Trp 5 and Trp 3–Met 34, which could function as a trigger for gating in response to a change in V_j. The released NTHs could then assemble into a plug that physically blocks the pore. The NTHs would not be released by the opposite potential, because they would be kept in position by their interaction with Met 34 (Supplementary Fig. 10). The release of any one of the six NTH would break down or destabilize the circular hydrogen bond network through the Asp 2–Thr 5 girdle, resulting in subconductance states of the channel. This would account for the report that the conformational change of a single subunit is sufficient to initiate V_j gating[37], although it is unclear whether the other five N termini adopt the same conformation as the one in action. In this way, the heteromeric oligomerization in a connexon would enable bipolar V_j gating[37], which allows the characteristic regulation of channel activity depending on the connexin isoforms expressed in each tissue.

The structure in this work could suggest a speculative V_j-gating model, in which the N termini have the chief role in sensing V_j within the conductive pore and in forming the plug to close the pore. This model is not the case for other voltage-sensitive ion channels containing the S4 helix as a voltage sensor[38], but is in accord with previous physiological studies[13,35,36]. However, we should consider an alternative possibility, because connexins are thought to use several gating mechanisms[39,40] and the previous electron microscopy structure was analysed in the condition that facilitates closure by chemical gating[12]. The C terminus of Cx26 is thought to be too short to form the gating particle suggested for Cx43 (ref. 41), but it is still associated with the chemical regulation of channel activity[42]. The structure in this work strongly suggests that the plug detected in the electron microscopy structure is composed of the assembly of Cx26 N termini. However, we do not rule out the possibility that the invisible cytoplasmic loop or the C terminus might contribute as a component. At present it is too premature to address the mechanism related to chemical gating from our structure.

Further discussions on the roles of the N terminus, the cytoplasmic loop and the C terminus are given in Supplementary Discussion.

METHODS SUMMARY

Human Cx26 was expressed in Sf9 insect cells using recombinant baculovirus. The gap junction channel was solubilized in dodecylmaltoside and purified sequentially by cation exchange and size-exclusion chromatography. Crystals were grown by the hanging-drop vapour diffusion method with PEG200 as a precipitant. The structure was determined by the single-isomorphous replacement combined with anomalous scattering (SIRAS) method, with phase extension by six-fold non-crystallographic (NCS) averaging.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 October 2008; accepted 9 February 2009.

1. Kumar, N. M. & Gilula, N. B. The gap junction communication channel. *Cell* **84,** 381–388 (1996).
2. Harris, A. L. Emerging issues of connexin channels: biophysics fills the gap. *Q. Rev. Biophys.* **34,** 325–472 (2001).
3. Foote, C. I., Zhou, L., Zhu, X. & Nicholson, B. J. The pattern of disulfide linkages in the extracellular loop regions of connexin 32 suggests a model for the docking interface of gap junctions. *J. Cell Biol.* **140,** 1187–1197 (1998).
4. Levin, M. Gap junctional communication in morphogenesis. *Prog. Biophys. Mol. Biol.* **94,** 186–206 (2007).
5. Saez, J. C., Berthoud, V. M., Branes, M. C., Martinez, A. D. & Beyer, E. C. Plasma membrane channels formed by connexins: their regulation and functions. *Physiol. Rev.* **83,** 1359–1400 (2003).
6. Kelsell, D. P., Dunlop, J. & Hodgins, M. B. Human diseases: clues to cracking the connexin code? *Trends Cell Biol.* **11,** 2–6 (2001).
7. Simon, A. M. & Goodenough, D. A. Diverse functions of vertebrate gap junctions. *Trends Cell Biol.* **8,** 477–483 (1998).
8. Unwin, P. N. & Zampighi, G. Structure of the junction between communicating cells. *Nature* **283,** 545–549 (1980).
9. Unwin, P. N. & Ennis, P. D. Two configurations of a channel-forming membrane protein. *Nature* **307,** 609–613 (1984).
10. Unger, V. M., Kumar, N. M., Gilula, N. B. & Yeager, M. Three-dimensional structure of a recombinant gap junction membrane channel. *Science* **283,** 1176–1180 (1999).
11. Fleishman, S. J., Unger, V. M., Yeager, M. & Ben-Tal, N. A C-α model for the transmembrane α helices of gap junction intercellular channels. *Mol. Cell* **15,** 879–888 (2004).
12. Oshima, A., Tani, K., Hiroaki, Y., Fujiyoshi, Y. & Sosinsky, G. E. Three-dimensional structure of a human connexin26 gap junction channel reveals a plug in the vestibule. *Proc. Natl Acad. Sci. USA* **104,** 10034–10039 (2007).
13. Harris, A. L., Spray, D. C. & Bennett, M. V. Kinetic properties of a voltage-dependent junctional conductance. *J. Gen. Physiol.* **77,** 95–117 (1981).
14. Verselis, V. K., Ginter, C. S. & Bargiello, T. A. Opposite voltage gating polarities of two closely related connexins. *Nature* **368,** 348–351 (1994).
15. Ebihara, L., Berthoud, V. M. & Beyer, E. C. Distinct behavior of connexin56 and connexin46 gap junctional channels can be predicted from the behavior of their hemi-gap-junctional channels. *Biophys. J.* **68,** 1796–1803 (1995).
16. Bukauskas, F. F., Bukauskiene, A., Bennett, M. V. & Verselis, V. K. Gating properties of gap junction channels assembled from connexin 43 and connexin 43 fused with green fluorescent protein. *Biophys. J.* **81,** 137–152 (2001).
17. Bukauskas, F. F. & Verselis, V. K. Gap junction channel gating. *Biochim. Biophys. Acta* **1662,** 42–60 (2004).
18. Muller, D. J., Hand, G. M., Engel, A. & Sosinsky, G. E. Conformational changes in surface structures of isolated connexin 26 gap junctions. *EMBO J.* **21,** 3598–3607 (2002).
19. Perkins, G. A., Goodenough, D. A. & Sosinsky, G. E. Formation of the gap junction intercellular channel requires a 30° rotation for interdigitating two apposing connexons. *J. Mol. Biol.* **277,** 171–177 (1998).
20. Skerrett, I. M. *et al.* Identification of amimo acid residues lining the pore of a gap junction channel. *J. Cell Biol.* **159,** 349–360 (2002).

21. Zhou, X. W. *et al.* Identification of a pore lining segment in gap junction hemichannels. *Biophys. J.* **72,** 1946–1953 (1997).
22. Kronengold, J., Trexler, E. B., Bukauskas, F. F., Bargiello, T. A. & Verselis, V. K. Single-channel SCAM identifies pore-lining residues in the first extracellular loop and first transmembrane domains of Cx46 hemichannels. *J. Gen. Physiol.* **122,** 389–405 (2003).
23. Suchyna, T. M., Xu, L. X., Gao, F., Fourtner, C. R. & Nicholson, B. J. Identification of a proline residue as a transduction element involved in voltage gating of gap junctions. *Nature* **365,** 847–849 (1993).
24. Laird, D. W. Life cycle of connexins in health and disease. *Biochem. J.* **394,** 527–543 (2006).
25. Sheridan, R. P., Levy, R. M. & Salemme, F. R. α-helix dipole model and electrostatic stabilization of 4-α-helical proteins. *Proc. Natl Acad. Sci. USA* **79,** 4545–4549 (1982).
26. Weber, P. A., Chang, H. C., Spaeth, K. E., Nitsche, J. M. & Nicholson, B. J. The permeability of gap junction channels to probes of different size is dependent on connexin composition and permeant-pore affinities. *Biophys. J.* **87,** 958–973 (2004).
27. Oh, S., Verselis, V. K. & Bargiello, T. A. Charges dispersed over the permeation pathway determine the charge selectivity and conductance of a Cx32 chimeric hemichannel. *J. Physiol. (Lond.)* **586,** 2445–2461 (2008).
28. Gong, X. Q. & Nicholson, B. J. Size selectivity between gap junction channels composed of different connexins. *Cell Commun. Adhes.* **8,** 187–192 (2001).
29. Trexler, E. B., Bukauskas, F. F., Kronengold, J., Bargiello, T. A. & Verselis, V. K. The first extracellular loop domain is a major determinant of charge selectivity in connexin46 channels. *Biophys. J.* **79,** 3036–3051 (2000).
30. Purnick, P. E., Benjamin, D. C., Verselis, V. K., Bargiello, T. A. & Dowd, T. L. Structure of the amino terminus of a gap junction protein. *Arch. Biochem. Biophys.* **381,** 181–190 (2000).
31. Miyazawa, A., Fujiyoshi, Y. & Unwin, N. Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **423,** 949–955 (2003).
32. Kelsell, D. P. *et al.* Connexin 26 mutations in hereditary non-syndromic sensorineural deafness. *Nature* **387,** 80–83 (1997).
33. Oshima, A., Doi, T., Mitsuoka, K., Maeda, S. & Fujiyoshi, Y. Roles of Met-34, Cys-64, and Arg-75 in the assembly of human connexin 26. Implication for key amino acid residues for channel formation and function. *J. Biol. Chem.* **278,** 1807–1816 (2003).
34. Oshima, A., Tani, K., Hiroaki, Y., Fujiyoshi, Y. & Sosinsky, G. E. Projection structure of a N-terminal deletion mutant of connexin 26 channel with decreased central pore density. *Cell Commun. Adhes.* **15,** 85–93 (2008).
35. Purnick, P. E., Oh, S., Abrams, C. K., Verselis, V. K. & Bargiello, T. A. Reversal of the gating polarity of gap junctions by negative charge substitutions in the N-terminus of connexin 32. *Biophys. J.* **79,** 2403–2415 (2000).
36. Oh, S., Rivkin, S., Tang, Q., Verselis, V. K. & Bargiello, T. A. Determinants of gating polarity of a connexin 32 hemichannel. *Biophys. J.* **87,** 912–928 (2004).
37. Oh, S., Abrams, C. K., Verselis, V. K. & Bargiello, T. A. Stoichiometry of transjunctional voltage-gating polarity reversal by a negative charge substitution in the amino terminus of a connexin 32 chimera. *J. Gen. Physiol.* **116,** 13–31 (2000).
38. Jan, L. Y. & Jan, Y. N. Structural elements involved in specific K+ channel functions. *Annu. Rev. Physiol.* **54,** 537–555 (1992).
39. Trexler, E. B., Bennett, M. V. L., Bargiello, T. A. & Verselis, V. K. Voltage gating and permeation in a gap junction hemichannel. *Proc. Natl Acad. Sci. USA* **93,** 5836–5841 (1996).
40. Peracchia, C. Chemical gating of gap junction channels; roles of calcium, pH and calmodulin. *Biochim. Biophys. Acta* **1662,** 61–80 (2004).
41. Delmar, M., Coombs, W., Sorgen, P., Duffy, H. S. & Taffet, S. M. Structural bases for the chemical regulation of connexin43 channels. *Cardiovasc. Res.* **62,** 268–275 (2004).
42. Tao, L. & Harris, A. L. 2-Aminoethoxydiphenyl borate directly inhibits channels composed of connexin26 and/or connexin32. *Mol. Pharmacol.* **71,** 570–579 (2007).
43. Baker, N. A., Sept, D., Joseph, S. & Holst, M. J. McCammon. J. A. Electrostatics of nanosystems: applications to microtubles and the ribosomes. *Proc. Natl Acad. Sci. USA* **98,** 10037–10041 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Tomizaki for help in the diffraction data collection on X06SA at the Swiss Light Source. This work was supported by Grants-in-Aid for Scientific Research (10687101, 16087206 and 18207006) and the GCOE program (A-041) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (to T.T.), the Japan Biological Informatics Consortium (to T.T.), the Strategic Japan-UK Cooperation Program of the Japan Science and Technology Agency (to T.T.), and Grants-in-Aid for Specially Promoted Research (to Y.F.) and the New Energy and Industrial Technology Development Organization (to Y.F.). We thank T. Walz for critical reading of this manuscript.

Author Contributions S.M., S.N., M.S., E.Y. and T.T. performed X-ray structural analysis. S.M., A.O., Y.F. and T.T. wrote the paper.

Author Information The atomic coordinate and the structure factor for the reported crystal structure have been deposited with the Protein Data Bank under accession code 2ZW3. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.T. (tsuki@protein.osaka-u.ac.jp).

METHODS

Expression and purification of Cx26. The human *Cx26* complementary DNA was amplified from a human liver cDNA library (Human liver QUICK-Clone cDNA, Clontech) by PCR and inserted via BamH I/EcoR I restriction sites into a pBlueBac4.5 (Invitrogen) baculovirus transfer vector. Recombinant baculovirus was made using the Bac-N-Blue system (Invitrogen). Baculovirus-infected Sf9 cells were grown at 27–28 °C and collected three days after infection. Purified Cx26 was obtained according to previously described and slightly modified methods[44]. In brief, collected cells were disrupted in alkali buffer containing 20 mM NaOH, 1 mM EDTA, 1 mM EGTA and 2 mM dithiothreitol (DTT), followed by ultracentrifugation to isolate the purified gap junction membrane fraction. The membrane fraction was then solubilized with 1–1.5% *n*-dodecyl-β-D-maltoside (DDM) in 10 mM CAPS (pH 10.5), 1 M NaCl and 10 mM DTT. The resulting supernatant was mixed with cation exchange resin, and Cx26 was eluted in 10 mM HEPES (pH 7.5), 0.01% DDM, 2 mM DTT and 500–1,000 mM NaCl. The protein was further purified by size-exclusion chromatography in 10 mM HEPES (pH 7.5), 200 mM NaCl, 2 mM DTT, 0.01% *n*-undecyl-β-D-maltoside (UDM), concentrated to 30 mg ml^{-1}, and used for crystallization. To prepare seleno-methionine (SeMet)-labelled protein[45], Sf9 cells were collected by centrifugation 24 h after infection, washed with sterilized PBS and transferred into medium devoid of methionine and supplemented with 20 mg l^{-1} SeMet and 150 mg l^{-1} L-cysteine. After a 4-h incubation, the cells were collected by centrifugation and transferred into medium supplemented with 50 mg l^{-1} SeMet and 150 mg l^{-1} L-cysteine. The cells were collected after two days, and SeMet-labelled protein was purified using the same protocol as for native protein.

Crystallization. Crystals were obtained by vapour diffusion (4 °C) by mixing equal volumes of protein solution and reservoir solution containing 100 mM potassium phosphate (pH 7.5), 100 mM KCl, 10 mM DTT, 0.5 mM EGTA and 16–18% PEG200. Crystals were dehydrated by gradually adding triethyleneglycol to a final concentration of 25–30% and flash frozen in liquid nitrogen. The Ta_6Br_{14} derivative was prepared by soaking the crystals in 1 mM Ta_6Br_{14} overnight.

X-ray data collection. Data sets were collected on BL44XU at SPring-8 with a DIP6040 imaging-plate detector (Bruker AXS). Two non-isomorphous native data sets, Native I and Native II, were collected at 3.5 Å and 4.0 Å, respectively. Isomorphous derivative crystals were generated for each native crystal (Derivative I, and Derivative II). Three data sets of tantalum derivative crystals were acquired by tuning X-rays at 0.9000 Å (remote), 1.2526 Å (peak), and 1.2552 Å (edge) (Derivative III). Diffraction data for the crystals, Native I, Native II and Derivative II were acquired with X-rays of 0.9000 Å, and that of Derivative I were acquired with X-rays of 1.2526 Å. Diffraction data of SeMet derivative crystals were collected with X-rays of 0.9000 Å (remote) and 0.9790 Å (edge). Another diffraction data set was acquired on the X06SA beamline at the Swiss Light Source, Paul Scherrer Institute, Villigen, Switzerland, using 1.7000 Å X-rays to detect anomalous dispersion effects of sulphur atoms in the native crystal using a Pilatus 6M detector. All X-ray experiments were performed at 100K. The SPring-8 diffraction data were processed and scaled with the Denzo, Scalepack[46] and CCP4 programs[47]. The SLS data were processed and scaled with the XDS and XSCALE programs[48]. The native crystals belonged to the space group *C*2 with cell dimensions of $a = 167.6$ Å, $b = 111.2$ Å, $c = 155.4$ Å and $\beta = 114.0$. Experimental conditions and statistics of intensity data acquisition are given in Supplementary Table 1.

Structure determination. Rotation function calculation of native crystals performed by POLARRFM[47] indicated a six-fold axis perpendicular to the crystallographic two-fold axis. The sites of the Ta_6Br_{14} clusters were determined in the difference Patterson map calculated with Native I data and Derivative III (remote) data at 6 Å resolution. Derivative III (peak) data were included in the phase determination, and anomalous dispersion effects of the heavy atoms were taken into account for the phase estimation. Assuming the tantalum cluster to be a single atom, the positional parameters and B-factor of the tantalum cluster were refined with the program SHARP[49]. Because of its large size, the tantalum cluster was effective for phase determination to no more than 6 Å resolution. The phases were refined and extended to 3.5 Å resolution by NCS averaging and solvent flattening using the program DM[50]. The preliminarily refined phase set was used to calculate a difference Fourier map of the SeMet derivative with coefficients of $[F_o(\text{remote}) - F_o(\text{edge})] \times \exp(i\alpha_c)$, in which α_c is the preliminarily refined phase. The SeMet sites were determined in the difference Fourier map. Of 42 methionine sites in the protein molecule, 36 were identified by selenium peaks higher than 4σ in the electron density distribution in the anomalous difference Fourier map. Thirty sites were used for phase calculations, whereas the other six sites were used to monitor the phase improvement steps.

Two native crystals, a tantalum derivative crystal, and a SeMet replacement crystal were used for the phase refinement by the multi-crystal averaging. Initial phases of each data set for the phase refinement were determined at 6 Å or 7 Å resolution. Those of the Native I crystal were determined by the SIRAS method using the Derivative I data at 6 Å resolution. Those of the Native II crystal equilibrated with 25% triethyleneglycol were determined by the single isomorphous replacement (SIR) method using the Derivative II data at 7 Å resolution. Those of the tantalum derivative crystal were determined by the Multiplewave anomalous dispersion (MAD) method using the Derivative III remote, peak, and edge data at 6 Å resolution. Those of the SeMet replacement crystal were determined by a method equivalent to the SIR method using the SeMet edge and remote data at 6 Å resolution.

The phase refinement was performed by multi-crystal averaging and six-fold NCS averaging combined with solvent flattening with the program DMMULTI[51]. The refinement procedure was monitored by *R* and *C* factors as a measure of consistency between observed and calculated structure factors, F_o and F_c, in which $R = \Sigma|F_o - F_c|/\ \Sigma\ |F_o|$, $C = \Sigma\ (F_o - < F_o >)(F_c - < F_c >)/\Sigma[(F_o - < F_o >)^2 (F_c - < F_c >)^2]^{1/2}$, and $< F_o >$ and $< F_c >$ are the averaged values of F_o and F_c in each resolution range. The phases were extended to 3.5 Å resolution, and the refinement converged well, with $R = 0.262$ and $C = 0.891$. An electron density map was calculated with the observed structure factors of Native I and the refined phases. The electron density map is called SIRAS/DM map in this paper.

Model building was performed using the programs O[52] and coot[53], and structural refinement was carried out under a tight restraint of non-crystallographic six-fold symmetry with the programs Crystallographic and NMR System (CNS)[54] and REFMAC[55]. The backbone of the protein was successfully traced in the SIRAS/DM map and in the composite omit map, in which aromatic residues are seen as bulky electron density (Supplementary Fig. 11a, b). To determine SeMet sites, a difference Fourier map was calculated at 6 Å resolution with coefficients of $[F_o(\text{remote}) - F_o(\text{edge})]\exp(i\alpha)$, in which F_o(remote) and F_o(edge) are the observed structure factors of the SeMet derivative measured by X-rays of 0.9000 Å and 0.9790 Å, and α is the phase of the Native I crystal determined by the SIR method with Ta derivative I combined with NCS averaging. The electron density of the Se atom at the N terminus was not detected in the difference map, probably due to disordered structure. To confirm the locations of the three disulphide bonds, which are close to each other, a native anomalous difference Fourier map was calculated at 4 Å resolution with the native F_o data acquired at the SLS and the phases calculated from a structural model refined by replacing their cysteine residues with alanine residues.

Crystallographic *R* and R_{free} for 5% of the reflections excluded from the refinement were calculated to monitor the structural refinement procedures. The close values of the final *R* and R_{free} were caused by the six-fold NCS restraints applied in the refinement[56]. The results of the structural analysis are summarized in Supplementary Table 1. The final *R* and R_{free} values were 33.7% and 35.1%, respectively. The main-chain dihedral angles for 84.0% of the non-glycine residues were in the most favoured region of the Ramachandran plot, 15.5% were in the allowed region, 0.5% were in the generously allowed region, and no residues were in the disallowed region. The refined structure was validated using the program PROCHECK[57]. All molecular graphics were created with Pymol[58].

44. Stauffer, K. A., Kumar, N. M., Gilula, N. B. & Unwin, N. Isolation and purification of gap junction channels. *J. Cell Biol.* **115,** 141–150 (1991).
45. Bellizzi, J. J. III, Widom, J., Kemp, C. W. & Clardy, J. Producing selenomethionine-labeled proteins with a baculovirus expression vector system. *Structure* **7,** R263–R267 (1999).
46. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
47. Collaborative Computational Project 4. The *CCP*4 suite: Programs for Protein Crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
48. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26,** 795–800 (1993).
49. Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Crystallogr. D* **59,** 2023–2030 (2003).
50. Cowtan, K. An automated procedure for phase improvement by density modification. *Joint CCP4 ESF-EACBM Newsletter Protein Crystallogr.* **31,** 34–38 (1994).
51. Cowtan, K. D. & Zhang, K. Y. Density modification for macromolecular phase improvement. *Prog. Biophys. Mol. Biol.* **72,** 245–270 (1999).
52. Jones, T. A., Zou, J. Y. & Cowan, S. W. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
53. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
54. Brunger, A. T. *et al.* Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).

55. Murshudov, G. N., Vagin, A. A. & Dadson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
56. Dodson, E., Kleywegt, G. J. & Wilson, K. Report of a workshop on the use of statistical validators in protein X-ray crystallography. *Acta. Crystallogr. D* **52,** 228–234 (1996).
57. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. *PROCHECK*: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26,** 283–291 (1993).
58. Delano, W. L. The *PyMOL Molecular Graphics System. v.0.99* (Delano Scientific, 2006).

LETTERS

Early assembly of the most massive galaxies

Chris A. Collins[1], John P. Stott[1], Matt Hilton[1,2,3], Scott T. Kay[4], S. Adam Stanford[5,6], Michael Davidson[7], Mark Hosmer[8], Ben Hoyle[9], Andrew Liddle[8], Ed Lloyd-Davies[8], Robert G. Mann[7], Nicola Mehrtens[8], Christopher J. Miller[10], Robert C. Nichol[9], A. Kathy Romer[8], Martin Sahlén[8], Pedro T. P. Viana[11,12] & Michael J. West[13]

The current consensus is that galaxies begin as small density fluctuations in the early Universe and grow by *in situ* star formation and hierarchical merging[1]. Stars begin to form relatively quickly in sub-galactic-sized building blocks called haloes which are subsequently assembled into galaxies. However, exactly when this assembly takes place is a matter of some debate[2,3]. Here we report that the stellar masses of brightest cluster galaxies, which are the most luminous objects emitting stellar light, some 9 billion years ago are not significantly different from their stellar masses today. Brightest cluster galaxies are almost fully assembled 4−5 billion years after the Big Bang, having grown to more than 90 per cent of their final stellar mass by this time. Our data conflict with the most recent galaxy formation models[4,5] based on the largest simulations of dark-matter halo development[1]. These models predict protracted formation of brightest cluster galaxies over a Hubble time, with only 22 per cent of the stellar mass assembled at the epoch probed by our sample. Our findings suggest a new picture in which brightest cluster galaxies experience an early period of rapid growth rather than prolonged hierarchical assembly.

Brightest cluster galaxies (BCGs) are located at the centres of galaxy clusters. They constitute a separate population from bright elliptical galaxies[6], and both their homogeneity and extreme luminosity have motivated their use as standard candles for cosmology[7–9]. Our investigation focuses on BCGs in the most distant X-ray-emitting galaxy clusters at redshifts of $z = 1.2$–1.5, where $1 + z$ is the expansion factor of the Universe relative to the present. It has been shown that X-ray cluster selection is currently the optimum strategy for an unbiased investigation of BCG evolution[10]. Properties of our BCGs and their host clusters are listed in Table 1. All five clusters were discovered serendipitously, and are the most distant clusters discovered in their respective X-ray surveys[11–15]. The cluster J2215 was discovered as part of the XMM Cluster Survey (XCS[16,17]) and has the highest redshift of any spectroscopically confirmed cluster[12,18].

The stellar mass of a BCG depends upon the hierarchical build-up of its host dark-matter halo and its stellar evolution history, along with the baryonic physics of the galaxy. We base our study of BCGs on photometry in the infrared wavebands J (1.26 μm) and K_s (2.14 μm). Infrared imaging is essential at these large distances to compensate for the redshifting of the early-type galaxy spectra. Also, these wavebands are less sensitive than optical light to the presence of young stars and are a more accurate tracer of the underlying old stellar population and, hence, of the stellar mass of the systems. Figure 1 shows an infrared image of the cluster J2235 from our sample (see also Supplementary Fig. 1).

We start by examining the ages of the stars themselves in these galaxies using the run of J–K_s colour evolution with redshift as shown in Fig. 2. For BCGs at the redshift of our sample the J–K_s colour predictions for the models separate clearly. For the comparison sample at lower redshift we use X-ray-selected clusters[19] which are well matched in mass to our own cluster sample. There is a remarkable agreement between the data and the hybrid model (see Fig. 2 legend), with all five BCGs lying within 0.05 magnitudes of their predicted colour, indicating a consistent epoch of formation for the majority of the constituent stars in all systems between redshifts $z_f = 3$–5, some 2–3 billion years (Gyr) after the Big Bang.

Table 1 | The properties of the host clusters and their BCGs

Cluster name	Redshift	X-ray luminosity (10^{44} erg s^{-1})	Cluster mass ($10^{14} M_\odot$)	BCG K_s (total)	J–K_s	Stellar mass ($10^{12} M_\odot$)
XLSS J022303.0−043622 (J0223)	1.22	$1.1^{+0.1}_{-0.1}$	1.0 ± 0.4	17.72 ± 0.01	1.82 ± 0.01	0.61 ± 0.08
XMMU J2235.3−2557 (J2235)	1.39	$11.4^{+0.7}_{-0.7}$	3.1 ± 0.7	17.34 ± 0.01	1.87 ± 0.02	1.26 ± 0.14
XMMXCS J2215.9−1738 (J2215)	1.46	$4.4^{+0.8}_{-0.6}$	1.8 ± 0.4	18.72 ± 0.01	1.83 ± 0.02	0.39 ± 0.05
RX J0848.9+4452 (J0849)	1.26	$3.3^{+0.9}_{-0.5}$	1.8 ± 0.4	17.00 ± 0.02	1.86 ± 0.03	1.30 ± 0.15
RDCS J1252.9−2927 (J1252)	1.24	$6.6^{+1.1}_{-1.1}$	2.6 ± 0.6	17.36 ± 0.03	1.83 ± 0.01	0.89 ± 0.11

The cluster X-ray luminosities are bolometric estimates taken from the literature and the cluster masses are M_{200} values (Supplementary Information). The errors on the cluster masses are based on the X-ray luminosity errors and the intrinsic uncertainty in the scaling relations. The J and K_s observations of J0223, J2235 and J2215 (Supplementary Fig. 1) were taken with the 8.2-m Subaru telescope and reach 5σ (s.d.) limiting magnitudes of $J \approx 23.7$ and $K_s \approx 22.8$ (23.1 in the case of J0223). The photometry for our data was calibrated using standard stars taken on the night in the Vega system. For comparison with previous observations we find that our J0223 BCG total K_s-band magnitude ($K_s = 17.72 \pm 0.01$) is in excellent agreement with the literature total magnitude[11] ($K_s = 17.76 \pm 0.04$, assuming a K_s-band conversion from AB to Vega system of −1.86). The photometry for J1252 and J0849 was sourced from the literature[14,21] and for these galaxies the total K_s magnitudes and J–K_s colours were measured in similar aperture sizes. All data have been analysed in an identical manner for direct comparison (see Supplementary Information). The errors on the stellar masses include all photometric errors and the uncertainty in the calibration with the semi-analytic model[4]. All errors are 1σ (s.d.). For each cluster we identified the brightest galaxy from the K_s-band magnitudes of all galaxies within 500 kpc of the cluster X-ray centroid because for approximately 95% of clusters the BCG lies within this radius[28]. All identified BCGs have optical spectra confirming their cluster membership[11,13–15,18].

[1]Astrophysics Research Institute, Liverpool John Moores University, Twelve Quays House, Egerton Wharf, Birkenhead, CH41 1LD, UK. [2]Astrophysics and Cosmology Research Unit, School of Mathematical Sciences, University of KwaZulu-Natal, Westville Campus, Private Bag X54001, Durban 4000, South Africa. [3]South African Astronomical Observatory, PO Box 9, Observatory, Cape Town 7935, South Africa. [4]Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, The University of Manchester, Manchester, M13 9PL, UK. [5]Department of Physics, University of California, Davis, California 95616, USA. [6]Institute of Geophysics and Planetary Physics, Lawrence Livermore National Laboratory, Livermore, California 94551, USA. [7]SUPA, Institute of Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, UK. [8]Astronomy Centre, University of Sussex, Falmer, Brighton, BN1 9QH, UK. [9]ICG, University of Portsmouth, Portsmouth, PO1 2EG, UK. [10]Cerro-Tololo Inter-American Observatory, National Optical Astronomy Observatory, 950 North Cherry Avenue, Tucson, Arizona 85719, USA. [11]Departmento de Matemática Aplicada da Faculdade de Ciências da Universiadade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal. [12]Centro de Astrofisica da Universidade do Porto, Rua das Estrelas, 4150-762 Porto, Portugal. [13]European Southern Observatory, Alonso de Córdova 3107, Vitacura, Casilla 19001, Santiago 19, Chile.

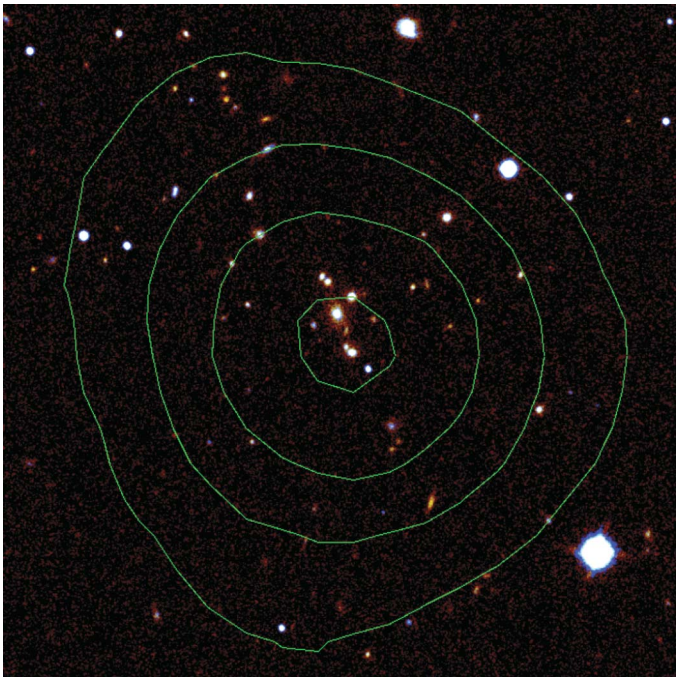


Figure 1 | Infrared image of the cluster J2235. An infrared image of the cluster J2235 at a redshift $z = 1.39$. Data were taken using the 8.2-m Subaru telescope. The image is combined from separate J and K_s exposures and shows the $1.5' \times 1.5'$ region surrounding the cluster centre. At this redshift $1.5'$ corresponds approximately to 0.75 Mpc. The green overlaid contours show the X-ray emission taken from the XMM-Newton XCS pipeline, smoothed with a Gaussian kernel. The X-ray peak coincides with the cluster centre and the position of the BCG. For a full description of the observations and data reduction see Supplementary Information.

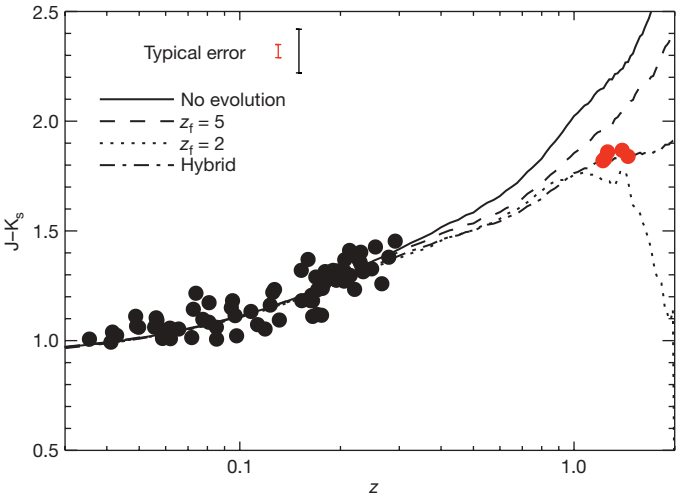


Figure 2 | The stellar evolution of BCGs with redshift. The J–K_s colour evolution for our five high-redshift BCGs (red) and 72 BCGs from the comparison sample[19] (black) which have host cluster masses in the same range as our high-redshift clusters and have available J and K_s photometry. The errors (s.d.) reported for the comparison sample[19] and our data are ~0.1 mag and ~0.02 mag respectively and are shown in the figure. This plot includes simple stellar population models[29] incorporating: no stellar evolution (solid line); passive evolution with formation epoch $z_f = 5$ (dashed line); passive evolution with formation epoch $z_f = 2$ (dotted line); a hybrid model with an exponentially decaying star-formation rate in which 50% of the BCG stellar content is formed by $z_f = 5$ and 80% by $z_f = 3$ (dot-dashed line), which is appropriate to the star-formation history predicted by the semi-analytic model[4]. The $z_f = 2$ and $z_f = 5$ stellar models are calculated assuming solar metallicity and a Salpeter initial mass function (IMF)[29], while the hybrid model was calculated with a Chabrier IMF[30]. The implied epoch of formation $z_f = 3$–5 (2–3 Gyr after the Big Bang) agrees well with other estimates of stellar ages determined for BCGs and early-type galaxies in clusters (see Supplementary Information). Throughout our analysis we assume a concordance cosmology of $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$; and $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, where Ω_Λ is the energy density associated with a cosmological constant. See the Supplementary Information for details of data reduction.

Turning our attention to the mass assembly of BCGs implied by our data, in Fig. 3 (see also Supplementary Table 1 and Supplementary Fig. 2) we show the estimates of stellar mass for our distant BCGs normalized to the average mass of the comparison sample at $z \leq 0.04$, which is $8.99\,(\pm 0.82) \times 10^{11}\ M_\odot$ (s.e.m.), where $M_\odot$ denotes the solar mass. Using a Tukey's biweight location estimator for robustness, for our five objects located at $z = 1.22$–1.46 we find an average stellar mass of $8.86\,(\pm 1.73) \times 10^{11} M_\odot$ (s.e.m.). The ratio of these estimates is 0.99 ± 0.21 (s.e.m.), indicating that on average the masses of the high-redshift BCGs are consistent with local counterparts.

To compare with theory we use the haloes from the Millennium Simulation[1] (http://www.mpa-garching.mpg.de/millennium) matched to the total mass of our clusters, estimated from their X-ray luminosity (see Supplementary Information). The mass range of our five clusters (Table 1) has excellent overlap with the combined $z = 1.08$ and $z = 1.5$ halo samples[4] (Supplementary Fig. 3). The predicted hierarchical mass build-up of BCGs in these 250 haloes is also shown in Fig. 3. The corresponding mass of the simulated BCGs has grown to an average of only $1.92\,(\pm 0.38) \times 10^{11} M_\odot$ (s.d.) by this time, some 22% of the observed value. The data are inconsistent with the prediction at the level of 4σ (one-tailed $P = 0.008$, degrees of freedom d.f. = 4; based on a Student's t distribution appropriate for small samples).

To check the stability of the BCG assembly predictions we selected massive haloes from the independent Durham semi-analytic model[5], which also uses the Millennium Simulation[1] but incorporates a different treatment of the baryon physics close to active galactic nuclei, partly to reproduce better the abundance of massive elliptical galaxies at high redshift. Using the same selection limits we find that the BCG mass fractions compared to the present day are $0.22^{+0.18}_{-0.09}$ at $z = 1$ and $0.17^{+0.12}_{-0.07}$ at $z = 1.5$, indicating good agreement between the two semi-analytical models.

It is well known that the estimates of stellar mass from photometry even for early-type galaxies such as BCGs depend on the underlying stellar evolution model used. To investigate this sensitivity we have applied three independent stellar population synthesis codes to early-type galaxies at the mean redshift of our sample ($z = 1.3$) using a range of model parameters (see Supplementary Table 1). These results show that the K_s band stellar mass estimates remain significantly different from the semi-analytic predictions (one-tailed $P \leq 0.02$, d.f. = 4) for the vast majority of parameters considered across the three models, reaching a value for one-tailed P of ≥ 0.05 in one of the three only if the stellar formation epoch z_f is less than 2.5 together with a stellar metallicity less than the solar value. This situation is incompatible with observations of BCGs and massive early-type galaxies in general (see Supplementary Information). We conclude that there remains a significant discrepancy between the recent semi-analytic models of galaxy formation coupled to the largest N-body simulations and the stellar masses of BCGs at the centres of the most massive clusters.

In comparison to recent studies[20], this work significantly extends the redshift baseline over which BCG evolution has been investigated to $z = 1.5$, equivalent to a look-back time of about 65% of the age of the Universe. Although the first glimpse of the $z > 1$ BCG population reveals galaxies with a range of stellar masses, there is on average considerably less stellar mass evolution than expected, with the bulk ($\geq 90\%$) of the stellar mass already in place by $z \approx 1.5$, corresponding to only about 4–5 Gyr after the Big Bang; the current models predict a considerably longer timescale of about 11 Gyr for the same growth, reaching 90% at $z \approx 0.2$.

Despite this, there is evidence that merging is still underway in our high-redshift sample. The BCG in J0849 at $z = 1.26$ has a nearby

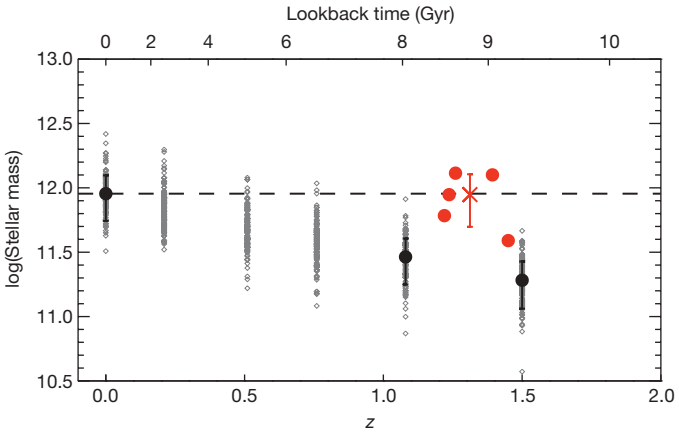


Figure 3 | The mass evolution of BCGs with redshift. The BCG mass estimates of our sample normalized to local galaxies at $z = 0.04$. The red cross is the estimated biweight location ($8.86 \times 10^{11} M_{\odot}$) and scale ($3.87 \times 10^{11} M_{\odot}$) of the sample. We calibrate the stellar masses by comparing the rest-frame absolute K_s magnitudes with the predicted magnitudes and corresponding stellar masses from the semi-analytic models[4]. This involves correcting the observed K_s values for: cosmological dimming; sampling different spectral regions of the galaxies resulting from the redshift (*k*-correction); and stellar evolution. The last two corrections are carried out using synthesized stellar spectra for early-type galaxies (appropriate to BCGs) from the hybrid stellar population model shown in Fig. 2. The *k*-correction is well understood over the wavelength range appropriate to our sample (0.9–2.2 μm), introducing an uncertainty of about 10% in the rest-frame absolute K_s magnitude estimates. The biweight scale provides a realistic estimate of the intrinsic error (s.d.) in the average mass using the hybrid model, however the total uncertainty in the inferred BCG mass is larger because it depends on the stellar evolution model used (see Supplementary Information). The grey diamonds show the individual BCG mass predictions[4] in 125 simulated clusters at each of six redshifts (0.0, 0.2, 0.5, 0.75, 1.08 and 1.5) above corresponding selection masses (4.7, 3.5, 2.8, 2.4, 1.5 and 1.0) in units of $10^{14} M_{\odot}$. The black-filled circles show the average value at each redshift (all errors are s.d.). The predictions are based on semi-analytic models of galaxy evolution. These use large *N*-body simulations such as the Millennium Simulation[1], which models the development of 2,160[3] cold dark-matter particles within a box that is over two billion light years per side. The semi-analytic techniques use the merger trees from the simulations and graft on analytical approximations to account for the complicated physics of the baryons in a range of ongoing processes associated with galaxy formation, such as: cooling, star formation, supernova outbursts and the growth of black holes in active galactic nuclei.

companion (projected separation of about 6 kiloparsecs, kpc) with which it is likely to undergo dissipationless merging in the future[21]. Of the other clusters in our sample, the BCG and its neighbour (projected separation of about 15 kpc) in J1252 are also possible merger candidates. Assuming that mergers take place in both these cases, the fraction of BCG stellar mass already assembled (based on the K_s fluxes of the main components) is ~84% and ~60% for J0849 and J1252 respectively, supporting the contention that most of the growth has actually already taken place in these two BCGs.

The timescale for the mass assemblage is similar to the age of the component stars (2–3 Gyr), a situation that appears to resemble classical monolithic collapse[22,23] rather than hierarchical formation. To form a galaxy of stellar mass $10^{12} M_{\odot}$ over 4 Gyr requires a mass deposition rate of about $250 M_{\odot}$ per year and an efficient mechanism of feeding the gas into the inner regions of the halo where it can form stars. Unfortunately, the merging process becomes inefficient for massive galaxies because merger-induced shocks lead to heating as opposed to radiative cooling of the gas[24]. One recent suggestion[25] is that the early assembly of massive galaxies at $z \geq 2$ is driven by narrow streams of dense cold gas which penetrate the shock-heated region, greatly increasing the efficiency of the gas deposition and associated star formation. Thus, in young BCGs the fraction of time the galaxy spends undergoing a major merger event could be less than 10%, with the stellar mass assembly dominated by this 'stream-fed' process[25]. Alternatively, a deficiency may lie in the semi-analytic treatment of the physical processes in the densest environments during early hierarchical assembly—a contention supported by the fact that current predictions are moderately consistent with observations of the evolution of luminous red galaxies[26,27], whereas our results, which focus on the most massive subset of this population, the BCGs, differ much more from the model predictions.

In a wider context, the hierarchical simulations and their semi-analytic prescriptions have arguably provided an excellent way of generating mock catalogues of galaxies to compare with real data, but our results show that they do not account for the assemblage history of all galaxies. Larger simulations may provide a better statistical probe of both the merging history of the largest haloes and cluster-mass trends. If BCGs collapsed and formed at high redshift in a single burst of intense star formation then they may well be dusty enough and in sufficient numbers to be detectable with the coming generation of submillimetre surveys, which will cover areas large enough to detect objects as rare as BCGs. The ongoing XCS survey will find many more high-redshift clusters and we anticipate that our results will stimulate independent studies of BCGs as new clusters are found in the redshift 'desert' beyond $z = 1.5$ from infrared and X-ray-based surveys such as eRosita.

Received 21 November 2008; accepted 9 February 2009.

1. Springel, V. *et al.* Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature* **435,** 629–636 (2005).
2. Kampakoglou, M., Trotta, R. & Silk, J. Monolithic or hierarchical star formation? A new statistical analysis. *Mon. Not. R. Astron. Soc.* **384,** 1414–1426 (2008).
3. van Dokkum, P. G. *et al.* Confirmation of the remarkable compactness of massive quiescent galaxies at z ≈ 2.3: early-type galaxies did not form in a simple monolithic collapse. *Astrophys. J. Lett.* **677,** L5–L8 (2008).
4. De Lucia, G. & Blaizot, J. The hierarchical formation of the brightest cluster galaxies. *Mon. Not. R. Astron. Soc.* **375,** 2–14 (2007).
5. Bower, R. G. *et al.* Breaking the hierarchy of galaxy formation. *Mon. Not. R. Astron. Soc.* **370,** 645–655 (2006).
6. Vale, A. & Ostriker, J. P. A non-parametric model for linking galaxy luminosity with halo/subhalo mass: are brightest cluster galaxies special? *Mon. Not. R. Astron. Soc.* **383,** 355–368 (2008).
7. Sandage, A. & Hardy, E. The redshift-distance relation. VIL absolute magnitudes of the first three ranked cluster galaxies as functions of cluster richness and Bautz-Morgan cluster type: the effect of q_o. *Astrophys. J.* **183,** 743–758 (1973).
8. Lauer, T. R. & Postman, M. The motion of the Local Group with respect to the 15,000 kilometer per second Abell cluster inertial frame. *Astrophys. J.* **425,** 418–438 (1994).
9. Collins, C. A. & Mann, R. G. The K-band Hubble diagram for brightest cluster galaxies in X-ray clusters. *Mon. Not. R. Astron. Soc.* **297,** 128–142 (1998).
10. Burke, D. J., Collins, C. A. & Mann, R. G. Cluster selection and the evolution of brightest cluster galaxies. *Astrophys. J. Lett.* **532,** L105–L108 (2000).
11. Bremer, M. N. *et al.* XMM-LSS discovery of a z = 1.22 galaxy cluster. *Mon. Not. R. Astron. Soc.* **371,** 1427–1434 (2006).
12. Stanford, S. A. *et al.* The XMM Cluster Survey: a massive galaxy cluster at z = 1.45. *Astrophys. J. Lett.* **646,** L13–L16 (2006).
13. Rosati, P. *et al.* An X-ray-selected galaxy cluster at Z = 1.26. *Astron. J.* **118,** 76–85 (1999).
14. Demarco, R. *et al.* VLT and ACS observations of RDCS J1252.9–2927: dynamical structure and galaxy populations in a massive cluster at z = 1.237. *Astrophys. J.* **663,** 164–182 (2007).
15. Mullis, C. R. *et al.* Discovery of an X-ray-luminous galaxy cluster at z = 1.4. *Astrophys. J. Lett.* **623,** L85–L88 (2005).
16. Romer, A. K., Viana, P. T. P., Liddle, A. R. & Mann, R. G. A serendipitous galaxy cluster survey with XMM: expected catalog properties and scientific applications. *Astrophys. J.* **547,** 594–608 (2001).
17. Sahlén, M. *et al.* The XMM Cluster Survey: forecasting cosmological and cluster scaling relation parameter constraints. Preprint at ⟨http://arxiv.org/abs/0802.4462⟩ (2008).
18. Hilton, M. *et al.* The XMM Cluster Survey: the dynamical state of XMMXCS J2215.9–1738 at z = 1.457. *Astrophys. J.* **670,** 1000–1009 (2007).
19. Stott, J. P., Edge, A. C., Smith, G. P., Swinbank, A. M. & Ebeling, H. Near-infrared evolution of brightest cluster galaxies in the most X-ray luminous clusters since z = 1. *Mon. Not. R. Astron. Soc.* **384,** 1502–1510 (2008).
20. Whiley, I. M. *et al.* The evolution of the brightest cluster galaxies since z ≈ 1 from the ESO Distant Cluster Survey (EDisCS). *Mon. Not. R. Astron. Soc.* **387,** 1253–1263 (2008).
21. Yamada, T. *et al.* Witnessing the hierarchical assembly of the brightest cluster galaxy in a cluster at z=1.26. *Astrophys. J. Lett.* **577,** L89–L92 (2002).
22. Eggen, O. J., Lynden-Bell, D. & Sandage, A. R. Evidence from the motions of old stars that the Galaxy collapsed. *Astrophys. J.* **136,** 748–767 (1962).

23. Larson, R. B. Dynamical models for the formation and evolution of spherical galaxies. *Mon. Not. R. Astron. Soc.* **166,** 585–616 (1974).
24. Binney, J. On the origin of the galaxy luminosity function. *Mon. Not. R. Astron. Soc.* **347,** 1093–1096 (2004).
25. Dekel, A. *et al.* Cold streams in early massive hot haloes as the main mode of galaxy formation. *Nature* **457,** 451–454 (2009).
26. Wake, D. A. *et al.* The 2df SDSS LRG and QSO survey: evolution of the luminosity function of luminous red galaxies to z = 0.6. *Mon. Not. R. Astron. Soc.* **372,** 537–550 (2006).
27. Almeida, C. *et al.* Luminous red galaxies in hierarchical cosmologies. *Mon. Not. R. Astron. Soc.* **386,** 2145–2160 (2008).
28. Lin, Y.-T. & Mohr, J. J. K-band properties of galaxy clusters and groups: brightest cluster galaxies and intracluster light. *Astrophys. J.* **617,** 879–895 (2004).
29. Bruzual, G. & Charlot, S. Stellar population synthesis at the resolution of 2003. *Mon. Not. R. Astron. Soc.* **344,** 1000–1028 (2003).
30. Chabrier, G. Galactic stellar and substellar initial mass function. *Publ. Astron. Soc. Pacif.* **115,** 763–795 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work is based in part on data collected at the Subaru Telescope, which is operated by the National Astronomical Observatory of Japan and the XMM-Newton, an ESA science mission funded by contributions from ESA member states and from NASA. We acknowledge financial support from Liverpool John Moores University and the STFC. M.H. acknowledges support from the South African National Research Foundation. IRAF is distributed by the National Optical Astronomy Observatories, which are operated by the Association of Universities for Research in Astronomy, Inc., under cooperative agreement with the National Science Foundation. We thank G. De Lucia for making simulation results available to us in tabular form, I. Tanaka for developing the MCSRED package used to reduce the MOIRCS data, M. Salaris for discussions on stellar population synthesis models and B. Maughan for discussions on cluster masses.

Author Contributions C.A.C. provided the scientific leadership, helped design the experiment, wrote the paper and led the interpretation. J.P.S. performed the photometry and data analysis and made major contributions to the interpretation. M.H. wrote the Subaru proposal, carried out the data reduction and photometric calibration, contributed to the analysis and interpretation and provided detailed comments on the manuscript. S.T.K. independently checked the cluster mass calculations. S.A.S. provided useful discussions on the data and comments on the manuscript. The remaining authors make up the team of the wider XCS project which led to the discovery of J2215. R.G.M., R.C.N., and A.K.R. made useful comments on the text.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.A.C. (cac@astro.livjm.ac.uk).

LETTERS

An anomalous positron abundance in cosmic rays with energies 1.5–100 GeV

O. Adriani[1,2], G. C. Barbarino[3,4], G. A. Bazilevskaya[5], R. Bellotti[6,7], M. Boezio[8], E. A. Bogomolov[9], L. Bonechi[1,2], M. Bongi[2], V. Bonvicini[8], S. Bottai[2], A. Bruno[6,7], F. Cafagna[7], D. Campana[4], P. Carlson[10], M. Casolino[11], G. Castellini[12], M. P. De Pascale[11,13], G. De Rosa[4], N. De Simone[11,13], V. Di Felice[11,13], A. M. Galper[14], L. Grishantseva[14], P. Hofverberg[10], S. V. Koldashov[14], S. Y. Krutkov[9], A. N. Kvashnin[5], A. Leonov[14], V. Malvezzi[11], L. Marcelli[11], W. Menn[15], V. V. Mikhailov[14], E. Mocchiutti[8], S. Orsi[10,11], G. Osteria[4], P. Papini[2], M. Pearce[16], P. Picozza[11,13], M. Ricci[17], S. B. Ricciarini[2], M. Simon[15], R. Sparvoli[11,13], P. Spillantini[1,2], Y. I. Stozhkov[5], A. Vacchi[8], E. Vannuccini[2], G. Vasilyev[9], S. A. Voronov[14], Y. T. Yurkin[14], G. Zampa[8], N. Zampa[8] & V. G. Zverev[14]

Antiparticles account for a small fraction of cosmic rays and are known to be produced in interactions between cosmic-ray nuclei and atoms in the interstellar medium[1], which is referred to as a 'secondary source'. Positrons might also originate in objects such as pulsars[2] and microquasars[3] or through dark matter annihilation[4], which would be 'primary sources'. Previous statistically limited measurements[5–7] of the ratio of positron and electron fluxes have been interpreted as evidence for a primary source for the positrons, as has an increase in the total electron+positron flux at energies between 300 and 600 GeV (ref. 8). Here we report a measurement of the positron fraction in the energy range 1.5–100 GeV. We find that the positron fraction increases sharply over much of that range, in a way that appears to be completely inconsistent with secondary sources. We therefore conclude that a primary source, be it an astrophysical object or dark matter annihilation, is necessary.

The results presented here are based on the data set collected by the PAMELA satellite-borne experiment[9] between July 2006 and February 2008. More than 10^9 triggers were accumulated during a total acquisition time of approximately 500 days. From these triggered events, 151,672 electrons and 9,430 positrons were identified in the energy interval 1.5–100 GeV. Results are presented as positron fraction—that is, the ratio of positron flux to the sum of electron and positron fluxes, $\phi(e^+)/(\phi(e^+)+\phi(e^-))$—and are shown in Table 1. The apparatus is a system of electronic particle detectors optimized for the study of antiparticles in the cosmic radiation (Supplementary Information section 1). It was launched from the Bajkonur cosmodrome on 15 June 2006 on board a satellite that was placed into a 70.0° inclination orbit, at an altitude varying between 350 km and 610 km. A permanent magnet spectrometer with a silicon tracking system allows the rigidity (momentum/charge, resulting in units of GV), and sign-of-charge of the incident particle to be determined. The interaction pattern in an imaging silicon-tungsten calorimeter allows electrons and positrons to be separated from protons.

The misidentification of protons is the largest source of background when estimating the positron fraction. This can occur if electron- and proton-like interaction patterns are confused in the calorimeter data. The proton-to-positron flux ratio increases from approximately 10^3 at 1 GV to approximately 10^4 at 100 GV. Robust positron identification is therefore required, and the residual proton background must be estimated accurately. The imaging calorimeter is 16.3 radiation lengths (0.6 nuclear interaction lengths) deep, so electrons and positrons develop well contained electromagnetic showers in the energy range of interest. In contrast, the majority of the protons will either pass through the calorimeter as minimum ionizing particles or interact deep in the calorimeter.

This is illustrated in Fig. 1, which shows $\mathcal{F}$, the fraction of calorimeter energy deposited inside a cylinder of radius 0.3 Molière radii, as a function of deflection (rigidity^{-1}). The axis of the cylinder is defined by extrapolating the particle track reconstructed in the spectrometer. For negatively-signed deflections, electrons are clearly visible as a horizontal band with $\mathcal{F}$ lying mostly between 0.4 and 0.7. For positively-signed deflections, the similar horizontal band is naturally associated with positrons, with the remaining points, mostly at $\mathcal{F} < 0.4$, designated as proton contamination (see Supplementary Information sections 2 and 3 for additional details concerning particle selection and background determination).

Figure 2 shows the positron fraction measured by the PAMELA experiment compared with other recent experimental data. The PAMELA data covers the energy range 1.5–100 GeV, with significantly higher statistics than other measurements. Two features are clearly visible in the data. At low energies (below 5 GeV) the PAMELA results are systematically lower than data collected during the 1990s, and at high energies (above 10 GeV) the PAMELA results show that the positron fraction increases significantly with energy.

Measurements of cosmic-ray positrons and electrons address a number of questions in contemporary astrophysics, such as the nature and distribution of particle sources in our Galaxy, and the subsequent propagation of cosmic rays through the Galaxy and the solar heliosphere. Positrons are believed to be mainly created in secondary production processes, that is, by the interaction of cosmic-ray nuclei with the interstellar gas. The solid line in Fig. 2 shows a calculation[1] based on such an assumption. Although this calculation is widely used, it does

[1]University of Florence, Department of Physics, Via Sansone 1, I-50019 Sesto Fiorentino, Florence, Italy. [2]INFN, Sezione di Florence, Via Sansone 1, I-50019 Sesto Fiorentino, Florence, Italy. [3]University of Naples "Federico II", Department of Physics, Via Cintia, I-80126 Naples, Italy. [4]INFN, Sezione di Naples, Via Cintia, I-80126 Naples, Italy. [5]Lebedev Physical Institute, Leninsky Prospekt 53, RU-119991 Moscow, Russia. [6]University of Bari, Department of Physics, Via Amendola 173, I-70126 Bari, Italy. [7]INFN, Sezione di Bari, Via Amendola 173, I-70126 Bari, Italy. [8]INFN, Sezione di Trieste, Padriciano 99, I-34012 Trieste, Italy. [9]Ioffe Physical Technical Institute, Polytekhnicheskaya 26, RU-194021 St Petersburg, Russia. [10]KTH, Department of Physics, AlbaNova University Centre, SE-10691 Stockholm, Sweden. [11]INFN, Sezione di Roma "Tor Vergata", Via della Ricerca Scientifica 1, I-00133 Rome, Italy. [12]IFAC, Via Madonna del Piano 10, I-50019 Sesto Fiorentino, Florence, Italy. [13]University of Rome "Tor Vergata", Department of Physics, Via della Ricerca Scientifica 1, I-00133 Rome, Italy. [14]Moscow Engineering and Physics Institute, Kashirskoe Shosse 31, RU-11540 Moscow, Russia. [15]Universität Siegen, D-57068 Siegen, Germany. [16]KTH, Department of Physics and The Oskar Klein Centre for Cosmoparticle Physics, AlbaNova University Centre, SE-10691 Stockholm, Sweden. [17]INFN, Laboratori Nazionali di Frascati, Via Enrico Fermi 40, I-00044 Frascati, Italy.

Table 1 | Summary of positron fraction results

Rigidity at spectrometer (GV)	Mean kinetic energy at top of payload (GeV)	Extrapolated $\frac{\phi(e^+)}{(\phi(e^+)+\phi(e^-))}$ at top of payload
1.5–1.8	1.64	$(0.0673^{+0.0014}_{-0.0013})$
1.8–2.2	1.99	(0.0607 ± 0.0012)
2.2–2.7	2.44	(0.0583 ± 0.0011)
2.7–3.3	2.99	(0.0551 ± 0.0012)
3.3–4.1	3.68	(0.0550 ± 0.0012)
4.1–5.0	4.52	(0.0502 ± 0.0014)
5.0–6.1	5.43	(0.0548 ± 0.0016)
6.1–7.4	6.83	(0.0483 ± 0.0018)
7.4–9.1	8.28	(0.0529 ± 0.0023)
9.1–11.2	10.17	$(0.0546^{+0.0029}_{-0.0028})$
11.2–15.0	13.11	$(0.0585^{+0.0030}_{-0.0031})$
15.0–20.0	17.52	$(0.0590^{+0.0040}_{-0.0041})$
20.0–28.0	24.02	(0.0746 ± 0.0059)
28.0–42.0	35.01	(0.0831 ± 0.0093)
42.0–65.0	53.52	$(0.106^{+0.022}_{-0.023})$
65.0–100.0	82.55	$(0.137^{+0.048}_{-0.043})$

The errors are one standard deviation. Details concerning particle selection and proton background determination can be found in Supplementary Information sections 2 and 3. The detection efficiencies for electrons and positrons are assumed to cancel, as the physical processes that these species undergo in the PAMELA detectors can be assumed to be identical across the energy range of interest. Possible bias arising from a sign-of-charge dependence on the acceptance due to the spectrometer magnetic field configuration and east–west effects caused by the Earth's magnetic field were excluded as follows. Effects due to the spectrometer magnetic field were studied using the PAMELA Collaboration's simulation software. No significant difference was found between the electron and positron detection efficiency above 1 GV. East–west effects, as well as contamination from re-entrant albedo particles (secondary particles produced by cosmic rays interacting with the Earth's atmosphere that are scattered upward but lack sufficient energy to leave the Earth's magnetic field and re-enter the atmosphere in the opposite hemisphere but at a similar magnetic latitude), are significant around and below the lowest permitted rigidity for a charged cosmic ray to reach the Earth from infinite distance, known as the geomagnetic cut-off. The geomagnetic cut-off for the PAMELA orbit varies from less than 100 MV for the highest orbital latitudes to ~15 GV for equatorial regions. In this work, only events with a measured rigidity exceeding the estimated vertical (PAMELA z-axis) geomagnetic cut-off by a factor of 1.3 were considered. This reduced east–west effects and re-entrant particle contamination to a negligible amount. The vertical geomagnetic cut-off was determined following the Størmer formalism on an event-by-event basis and using orbital parameters reconstructed at a rate of 1 Hz.

not account for uncertainties related to the production of secondary positrons and electrons (see ref. 10). Uncertainties arise because of incomplete knowledge of (1) the primary cosmic-ray nuclei spectra, (2) modelling of interaction cross-sections, (3) modelling of cosmic-ray propagation in the Galaxy and (4) solar modulation effects.

The low energy data from previous experiments (CAPRICE94[11], HEAT95[6] and AMS-01[12]) match the calculated secondary fraction while the PAMELA data are clearly lower. This points to charge-sign-dependent solar modulation effects. The solar wind modifies the energy spectra of cosmic rays within the Solar System. This effect is called solar modulation, and has a significant effect on cosmic rays with energies less than about 10 GeV. The amount of solar modulation depends on solar activity, which has an approximately sinusoidal time dependence and is most evident at solar maximum, when the low-energy cosmic-ray flux is at a minimum. The peak-to-peak period is 11 years, but a complete 'solar cycle' is 22 years long because at each maximum the polarity of the solar magnetic field reverses. The low energy difference between the PAMELA and other, older, results can be interpreted as a consequence of charge dependent solar modulation effects (Supplementary Information section 4). These older results were collected during the previous polarity of the solar cycle. A balloon-borne experiment which flew in June 2006 has also observed a suppressed positron fraction[13] at low energies, but with large statistical uncertainties.

Above 5 GeV, the PAMELA positron fraction agrees with the most recent measurements[5–7]. Although too statistically limited to draw any significant conclusions, these high energy measurements indicate a flatter positron fraction than expected from secondary production models. Now, PAMELA data clearly show that the positron fraction increases significantly with energy. Besides the uncertainties previously discussed, those on the primary electron spectrum are also relevant. The electron injection spectrum at source is expected to have a power law index of approximately −2 (ref. 14) and be equal to that of protons[15] up to about 1 TeV. When the energy losses of

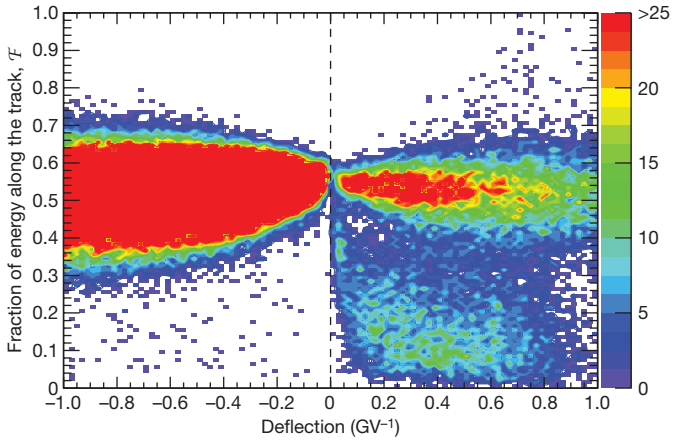


Figure 1 | Calorimeter energy fraction, $\mathcal{F}$. The fraction of calorimeter energy deposited inside a cylinder of radius 0.3 Molière radii, as a function of deflection. The number of events per bin is shown in different colours, as indicated in the colour scale. The axis of the cylinder is defined by extrapolating the particle track reconstructed by the spectrometer. The Molière radius is an important quantity in calorimetry, as it quantifies the lateral spread of an electromagnetic shower (about 90% of the shower energy is contained in a cylinder with a radius equal to 1 Molière radius), and depends only on the absorbing material (tungsten in this case). The events were selected requiring a match between the momentum measured by the tracking system and the total detected energy and requiring that the electromagnetic shower starts developing in the first planes of the calorimeter. The particle identification was tuned to reject 99.9% of the protons, while selecting >95% of the electrons or positrons.

primary cosmic rays during their propagation are taken into account, electrons are expected to have a harder spectrum than positrons if these are mostly of secondary origin. Hence, the positron fraction is expected to fall as a smooth function of increasing energy. Therefore, PAMELA positron fraction data cannot be understood by standard models describing the secondary production of cosmic rays. Either a significant modification in the acceleration and propagation models for cosmic rays is needed, or a primary component is present (for more details, see ref. 16). There are several interesting candidates for a primary component, including the annihilation of dark matter particles in the vicinity of our Galaxy and a contribution from nearby astrophysical sources, such as pulsars or microquasars.

The energy budget of the Universe can be broken down into baryonic matter (about 5%), dark matter (about 23%) and dark energy (about 72%)[17]. Many particle candidates have been proposed for the dark matter component. The most widely studied are weakly interacting massive particles (WIMPs), such as the neutralino from supersymmetric models[4] and the lightest Kaluza Klein particle from extra dimension models[18,19]. High energy antiparticles such as positrons and antiprotons (see ref. 20 and references within) can be produced during the annihilation or decay of these dark matter particles in our Galaxy. In a previous publication[21], we presented the antiproton-to-proton flux ratio in the energy range 1–100 GeV. The data follow the trend expected from secondary production calculations for antiprotons. Therefore, if the PAMELA positron results have a component due to dark matter this has to annihilate or decay into mostly leptonic final states. Furthermore, heavy WIMP candidates or large boost factors (see, for example, refs 22, 23) associated with non-uniform clumps in the dark matter distribution are required. It is worth pointing out that our antiproton-to-proton flux ratio data[21] limit significantly the boost factor for thermal WIMP candidates (ref. 24). WIMPs of non-thermal origin[25] can also be considered as explanations for both PAMELA positron and antiproton results. This model predicts a sharp decrease in the primary positron spectrum above 100 GeV, an energy range that PAMELA is exploring and will be soon able to clarify.

The possible production of positrons from nearby astrophysical sources, such as pulsars[2,26,27] and microquasars[3], must be taken into

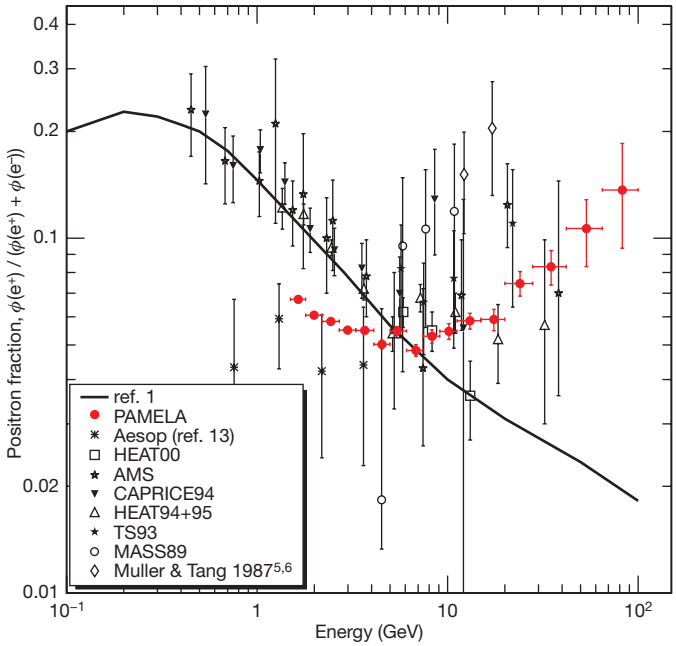


Figure 2 | PAMELA positron fraction with other experimental data and with secondary production model. The positron fraction measured by the PAMELA experiment compared with other recent experimental data (see refs 5–7, 11–13, 30, and references within). The solid line shows a calculation[1] for pure secondary production of positrons during the propagation of cosmic rays in the Galaxy without reacceleration processes. Error bars show 1 s.d.; if not visible, they lie inside the data points.

account when interpreting potential dark matter signals. A pulsar magnetosphere is a well known cosmic particle accelerator. The details of the acceleration processes are as yet unclear, but electrons are expected to be accelerated in the magnetosphere, where they induce an electromagnetic cascade. This process results in electrons and positrons that can escape into the interstellar medium, contributing to the cosmic-ray electron and positron components. As the energy spectrum of these particles is expected to be harder than that of the secondary positrons, such pulsar-originated positrons may dominate the high energy end of the cosmic-ray positron spectrum. But because of the energy losses of electrons and positrons during their propagation, just one or a few nearby pulsars can contribute significantly to the positron energy spectrum (see, for example, refs 28, 29).

The PAMELA positron data presented here are insufficient to distinguish between astrophysical primary sources and dark matter annihilation. However, PAMELA will soon present results concerning the energy spectra of primary cosmic rays—such as electrons, protons and higher mass nuclei—that will significantly constrain the secondary production models, thereby lessening the uncertainties on the high energy behaviour of the positron fraction. Furthermore, the experiment is continuously taking data and the increased statistics will allow the measurement of the positron fraction to be extended up to an energy of about 300 GeV. The combination of these efforts will help in discriminating between various dark matter and pulsar models put forward to explain both our results and the ATIC[8] results. New important information will soon come also from the FERMI satellite that is studying the diffuse Galactic cosmic γ-ray spectrum. Pulsars are predominantly distributed along the Galactic plane, while dark matter is expected to be spherically distributed as an extended halo and highly concentrated at the Galactic Centre. The diffuse γ-ray spectrum is sensitive to these different geometries. Furthermore, PAMELA is measuring the energy spectra of both electrons (up to ~500 GeV) and positrons (up to ~300 GeV). These data will clarify if the ATIC results[8] are due to a significantly large component of pair-produced electrons and positrons (to explain the high energy ATIC data, the positron fraction should exceed 0.3 above 300 GeV), hence pointing to primary positron sources, or to a hardening of the electron spectrum with a more mundane explanation.

Received 28 October 2008; accepted 6 February 2009.

1. Moskalenko, I. V. & Strong, A. W. Production and propagation of cosmic-ray positrons and electrons. *Astrophys. J.* **493,** 694–707 (1998).
2. Atoian, A. M., Aharonian, F. A. & Volk, H. J. Electrons and positrons in the galactic cosmic rays. *Phys. Rev. D* **52,** 3265–3275 (1995).
3. Heinz, S. & Sunyaev, R. Cosmic rays from microquasars: A narrow component in the CR spectrum. *Astron. Astrophys.* **390,** 751–766 (2002).
4. Jungman, G., Kamionkowski, M. & Griest, K. Supersymmetric dark matter. *Phys. Rep.* **267,** 195–373 (1996).
5. Golden, R. L. *et al.* Measurement of the positron to electron ratio in the cosmic rays above 5 GeV. *Astrophys. J.* **457,** L103–L106 (1996).
6. Barwick, S. W. *et al.* Measurements of the cosmic-ray positron fraction from 1 to 50 GeV. *Astrophys. J.* **482,** L191–L194 (1997).
7. Aguilar, M. *et al.* Cosmic-ray positron fraction measurement from 1 to 30 GeV with AMS-01. *Phys. Lett. B* **646,** 145–154 (2007).
8. Chang, J. *et al.* An excess of cosmic ray electrons at energies of 300–800 GeV. *Nature* **456,** 362–365 (2008).
9. Picozza, P. *et al.* PAMELA — A payload for antimatter matter exploration and light-nuclei astrophysics. *Astropart. Phys.* **27,** 296–315 (2007).
10. Delahaye, T. *et al.* Galactic secondary positron flux at the Earth. Preprint at ⟨http://arXiv.org/abs/0809.5268v3⟩ (2008).
11. Boezio, M. *et al.* The cosmic-ray electron and positron spectra measured at 1 AU during solar minimum activity. *Astrophys. J.* **532,** 653–669 (2000).
12. Alcaraz, J. *et al.* Leptons in near earth orbit. *Phys. Lett. B* **484,** 10–22 (2000).
13. Clem, J. & Evenson, P. in *Proc. 30th Intl Cosmic Ray Conf.* Vol. 1 (eds Caballero, R. *et al.*) 477–480 (Universidad Nacional Autónoma de México, 2008).
14. Aharonian, F. *et al.* First detection of a VHE gamma-ray spectral maximum from a cosmic source: HESS discovery of the Vela X nebula. *Astron. Astrophys.* **448,** L43–L47 (2006).
15. Berezhko, E. G., Ksenofontov, L. T. & Völk, H. J. Emission of SN 1006 produced by accelerated cosmic rays. *Astron. Astrophys.* **395,** 943–953 (2002).
16. Serpico, P. On the possible causes of a rise with energy of the cosmic ray positron fraction. *Phys. Rev. D* **79,** 021302 (2009).
17. Komatsu, E. *et al.* Five-year Wilkinson microwave anisotropy probe observations: Cosmological interpretation. *Astrophys. J. Suppl. Ser.* **180,** 330–376 (2009).
18. Servant, G. & Tait, T. M. P. Is the lightest Kaluza-Klein particle a viable dark matter candidate? *Nucl. Phys. B* **650,** 391–419 (2003).
19. Cheng, H. C., Feng, J. L. & Matchev, K. T. Kaluza-Klein dark matter. *Phys. Rev. Lett.* **89,** 211301 (2002).
20. Bertone, G., Hooper, D. & Silk, J. Particle dark matter: Evidence, candidates and constraints. *Phys. Rep.* **405,** 279–390 (2005).
21. Adriani, O. *et al.* A new measurement of the antiproton-to-proton flux ratio up to 100 GeV in the cosmic radiation. *Phys. Rev. Lett.* **102,** 051101 (2009).
22. Cholis, I., Dobler, G., Finkbeiner, D. P., Goodenough, L. & Weiner, N. The case for a 700+ GeV WIMP: Cosmic ray spectra from ATIC and PAMELA. Preprint at ⟨http://arXiv.org/abs/0811.3641v1⟩ (2008).
23. Bergström, L., Bringmann, T. & Edsjö, J. New positron spectral features from supersymmetric dark matter: A way to explain the PAMELA data? *Phys. Rev. D* **78,** 103520 (2008).
24. Donato, F., Maurin, D., Brun, P., Delahaye, T. & Salati, P. Constraints on WIMP dark matter from the high energy PAMELA $\bar{p}/p$ data. *Phys. Rev. Lett.* **102,** 071301 (2009).
25. Grajek, P., Kane, G., Phalen, D. J., Pierce, A., &. and Watson, A. Is the PAMELA positron excess winos? Preprint at ⟨http://arXiv.org/abs/0812.4555v1⟩ (2008).
26. Grimani, C. Pulsar birthrate set by cosmic-ray positron observations. *Astron. Astrophys.* **418,** 649–653 (2004).
27. Büsching, I., de Jager, O. C., Potgieter, M. S. & Venter, C. A cosmic-ray positron anisotropy due to two middle-aged, nearby pulsars? *Astrophys. J.* **78,** L39–L42 (2008).
28. Yuksel, H., Kistler, M. D. & Stanev, T. TeV gamma rays from Geminga and the origin of the GeV positron excess. Preprint at ⟨http://arXiv.org/abs/0810.2784v2⟩ (2008).
29. Hooper, D., Blasi P. & Serpico, P. D. Pulsars as the sources of high energy cosmic ray positrons. *J. Cosmol. Astropart. Phys.* **01,** 025 (2009).
30. Beatty, J. J. *et al.* New measurement of the cosmic-ray positron fraction from 5 to 15GeV. *Phys. Rev. Lett.* **93,** 241102 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Marinucci for discussions concerning statistical methods, D. Müller, S. Swordy and their group at University of Chicago, G. Bellettini and G. Chiarelli for discussions about the data analysis and L. Bergström for comments on the interpretation of our results. We acknowledge support from The Italian Space Agency (ASI), Deutsches Zentrum für Luftund Raumfahrt (DLR), The Swedish National Space Board, The Swedish Research Council, The Russian Space Agency (Roscosmos) and The Russian Foundation for Basic Research.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.P. (Piergiorgio.Picozza@roma2.infn.it).

LETTERS

Emergence of the persistent spin helix in semiconductor quantum wells

J. D. Koralek[1], C. P. Weber[1,2], J. Orenstein[1,3], B. A. Bernevig[4], Shou-Cheng Zhang[5], S. Mack[6] & D. D. Awschalom[6]

According to Noether's theorem[1], for every symmetry in nature there is a corresponding conservation law. For example, invariance with respect to spatial translation corresponds to conservation of momentum. In another well-known example, invariance with respect to rotation of the electron's spin, or SU(2) symmetry, leads to conservation of spin polarization. For electrons in a solid, this symmetry is ordinarily broken by spin–orbit coupling, allowing spin angular momentum to flow to orbital angular momentum. However, it has recently been predicted that SU(2) can be achieved in a two-dimensional electron gas, despite the presence of spin–orbit coupling[2]. The corresponding conserved quantities include the amplitude and phase of a helical spin density wave termed the 'persistent spin helix'[2]. SU(2) is realized, in principle, when the strengths of two dominant spin–orbit interactions, the Rashba[3] (strength parameterized by α) and linear Dresselhaus[4] (β_1) interactions, are equal. This symmetry is predicted to be robust against all forms of spin-independent scattering, including electron–electron interactions, but is broken by the cubic Dresselhaus term (β_3) and spin-dependent scattering. When these terms are negligible, the distance over which spin information can propagate is predicted to diverge as α approaches β_1. Here we report experimental observation of the emergence of the persistent spin helix in GaAs quantum wells by independently tuning α and β_1. Using transient spin-grating spectroscopy[5], we find a spin-lifetime enhancement of two orders of magnitude near the symmetry point. Excellent quantitative agreement with theory across a wide range of sample parameters allows us to obtain an absolute measure of all relevant spin–orbit terms, identifying β_3 as the main SU(2)-violating term in our samples. The tunable suppression of spin relaxation demonstrated in this work is well suited for application to spintronics[6,7].

Transient spin-grating spectroscopy (TSG) is a powerful tool for searching for the persistent spin helix (PSH) because it enables measurement of the lifetime of spin polarization waves as a function of wavevector, **q**. In TSG, spin polarization waves of well-defined **q** are generated by exciting a two-dimensional electron gas (2DEG) with two non-collinear beams of light from a femtosecond laser. When the two incident pulses of light are linearly polarized in orthogonal directions, interference generates stripes of alternating photon helicity in the sample. Because of the optical orientation[8] effect in III–V semiconductors, the photon helicity wave generates a spin polarization wave in the 2DEG. The wavevector is varied by changing the angle between the interfering beams. The spin wave imprinted in the 2DEG acts as an optical diffraction grating, allowing its subsequent temporal evolution to be monitored by the diffraction of a time-delayed probe pulse[9].

In Fig. 1a we show a set of TSG decay curves for a 2DEG in an asymmetrically modulation-doped GaAs quantum well, which is expected to have both Rashba and Dresselhaus spin–orbit interactions. Each curve represents the decay of a spin grating at a specific **q**. The decay at $q \equiv |\mathbf{q}| = 0$ (not shown), measured by time-resolved Faraday rotation[10], follows a single exponential over nearly three orders of magnitude. With increasing **q**, the decay evolves towards the sum of two exponentials with nearly equal weights but very different rate constants. We fit the TSG decay curves with double exponentials and plot the resulting spin lifetimes in Fig. 1b, immediately observing very unusual spin-diffusion properties. The rapidly decaying component of

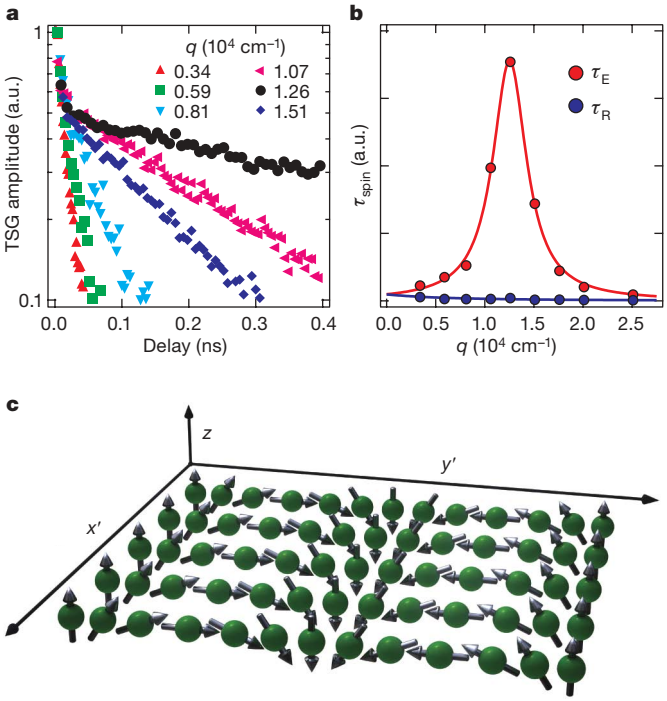


Figure 1 | Double-exponential decay of transient spin gratings. a, TSG decay curves at various wavevectors, **q**, for an asymmetrically doped 2DEG with a mixture of Rashba and Dresselhaus spin–orbit couplings. a.u., arbitrary units. **b**, Lifetimes for the spin–orbit-enhanced (τ_E) and -reduced (τ_R) helix modes extracted from double-exponential fits to the data in **a**. The solid lines are a theoretical fit (see text) using a single set of spin–orbit parameters for both helix modes. Error bars (s.d.) are the size of the data points. **c**, Illustration of a helical spin wave, which is one of the normal modes of the spin–orbit coupled 2DEG. In this picture, z is the growth direction [001], and the axes x' and y' respectively refer to the [11] and [1$\bar{1}$] directions in the plane of the 2DEG. The green spheres represent electrons whose spin directions are given by the arrows.

[1]Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [2]Department of Physics, Santa Clara University, Santa Clara, California 95053, USA. [3]Department of Physics, University of California, Berkeley, California 94720, USA. [4]Princeton Center for Theoretical Science, Princeton University, Princeton, New Jersey 08540, USA. [5]Department of Physics, Stanford University, Stanford, California 94305, USA. [6]Center for Spintronics and Quantum Computation, University of California, Santa Barbara, California 93106, USA.

the TSG (lifetime, τ_R) displays ordinary diffusion in the sense that the spin lifetime is peaked at $q = 0$. On the other hand, the lifetime of the slowly decaying component (τ_E), is peaked sharply at a non-zero value of q.

The salient features of Fig. 1a, b were predicted by recent quantitative theories of spin propagation in a 2DEG in the presence of spin–orbit coupling. The effects on spin propagation of the Rashba interaction term in the Hamiltonian, $H_R = \alpha\hbar v_F(k_y\sigma_x - k_x\sigma_y)$, where v_F is the Fermi velocity, $\hbar\mathbf{k} \equiv \hbar(k_x, k_y, k_z)$ is the electron momentum ($\hbar$ denoting Planck's constant divided by 2π) and σ_x and σ_y are Pauli matrices, were studied in refs 11, 12. The term H_R corresponds to an in-plane, $\mathbf{k}$-dependent magnetic field, $\mathbf{b}_R = \alpha\hbar v_F(k_y\hat{\mathbf{x}} - k_x\hat{\mathbf{y}})$, that leads to precession of the electron's spin. It was found that, in the presence of $\mathbf{b}_R$, the normal modes of the system are helical waves of spin polarization in which the spin direction rotates in the plane normal to the 2DEG and parallel to the wavevector, $\mathbf{q}$ (Fig. 1c). For each $\mathbf{q}$, there are two helical modes with opposite senses of rotation. The lifetime of the mode whose sense of rotation matches the precession of the electron's spin is enhanced, and the lifetime of the other is reduced. A striking prediction is that, for a range of $\mathbf{q}$ values, the spin–orbit-enhanced lifetime will exceed that of a uniform ($q = 0$) spin polarization. This contrasts with ordinary diffusion, in which the decay rate for spin excitations scales as q^2 and the spin lifetime is always greatest at $q = 0$. (The same conclusions are reached when the linear Dresselhaus term, $H_D = \beta_1\hbar v_F(k_x\sigma_x - k_y\sigma_y)$, is assumed to be the only spin–orbit interaction). Experimental support for these predictions was reported[13], in which a maximum spin lifetime at non-zero q was observed in nominally symmetric GaAs quantum wells, where H_D dominates the spin–orbit Hamiltonian.

Recently, this theory has been extended to predict the lifetimes of helix modes in the presence of both H_R and H_D (ref. 2). In particular, it was predicted that as the two couplings approach equal strength, the spin–orbit-enhanced mode evolves to the PSH; that is, the lifetime tends to infinity for $\mathbf{q} = \mathbf{q}_{PSH}$. As discussed above, the stability of the PSH is a manifestation of SU(2) symmetry at this point in parameter space. Although conservation of the x' component (Fig. 1c) of spin, or U(1) symmetry, was noted previously[14], SU(2) symmetry implies conservation of the amplitude and phase of the PSH as well. The theory also predicts quantitatively how the persistence of the helix degrades with detuning from the SU(2) point, by variation either of $\mathbf{q}$ or the α/β_1 ratio. The theory has been extended further[15] by the inclusion of the SU(2)-breaking effects of the cubic Dresselhaus coupling

$$H_{cD} = -4\beta_3\frac{\hbar v_F}{k_F^2}(k_x k_y^2\sigma_x - k_y k_x^2\sigma_y)$$

(k_F is the Fermi wavevector) which is always present at some level because of the non-zero width of the quantum well (see below).

The predictions of helical spin modes described above are clearly evident in the TSG results shown in Fig. 1. The initial condition created by the two pump pulses—a sinusoidal of variation S_z at $t = 0$—is equivalent to two equal-amplitude S_y–S_z helices of opposite pitch. Each of these normal modes then decays independently with its own characteristic decay rate, corresponding to the spin–orbit-enhanced and -reduced helix lifetimes (τ_E and τ_R, respectively). The reduced lifetimes shown in Fig. 1b peak at $q = 0$, whereas the enhanced lifetime is greatest at a finite value of $\mathbf{q}$. The solid lines are a fit to the theory of ref. 15 using a single set of spin–orbit parameters for both the enhanced and reduced helix modes.

The fact that the dispersion of both branches is accurately fitted by a single set of spin–orbit parameters suggests that spin helices are indeed the normal modes of our spin–orbit-coupled 2DEGs. The theoretical fits provide us with values for α, β_1, β_3 and D_S (the spin-diffusion coefficient), which we then use to guide us in engineering quantum wells with the longest spin-helix lifetimes. To tune the spin–orbit Hamiltonian, we have designed a series of quantum-well samples with varying doping asymmetries and well widths.

Figure 2 summarizes the spin–orbit tuning results. To tune the Rashba interaction, which arises from asymmetry in the electron's confinement potential, we varied the relative concentration of remote dopants on the two sides of the 2DEG (keeping the total dopant concentration fixed). These measurements were performed at $T = 75$ K, at which temperature the enhanced-mode lifetimes are greatest (see discussion of T dependence below). The enhanced spin lifetimes, τ_E, are plotted as functions of q in Fig. 2a, for a set of 12-nm-wide quantum wells with varying amounts of doping asymmetry. The maximum lifetime and the wavevector at which it occurs grow monotonically with increasing dopant asymmetry. Figure 2b shows the spin–orbit parameters extracted from comparison of the dispersion curves with the theory of ref. 15, for each of the samples. The parameters α, β_1 and β_3 are plotted as functions of normalized asymmetry, defined as the difference between the concentrations of dopant ions on either side of the well, divided by the total dopant concentration. The variation in α is well approximated by a straight line that extrapolates to zero coupling as the asymmetry parameter goes to zero. The data for the nominally symmetric sample display a residual Rashba coupling, which we attribute to the inherent asymmetry in the growth of the quantum wells[13,16–18]. Although only τ_E is shown in Fig. 2a, the high decay rates are accurately described by the same set of parameters.

The linear Dresselhaus interaction is related to the degree of confinement of the electrons (that is, it is proportional to $\langle k_z^2\rangle$). We tune the linear Dresselhaus interaction by varying the width, d, of each quantum well, with the normalized asymmetry fixed at unity. Experimentally determined spin lifetimes and theoretical fits are plotted in Fig. 2c for values of d ranging from 7 to 15 nm. The peak value of τ_E shows a clear maximum, suggesting that as d is varied the spin–orbit Hamiltonian approaches and then recedes from the SU(2) point. The curves generated from the theory of ref. 15 fit the data very

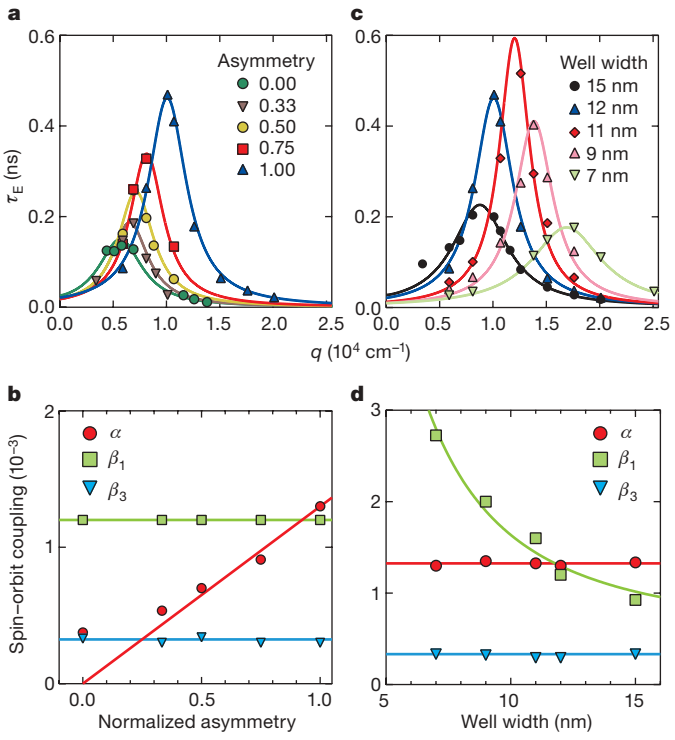


Figure 2 | Rashba and linear Dresselhaus tuning. a, **c**, Lifetimes of the enhanced helix mode are shown for samples with varying degrees of doping asymmetry (**a**) and well width (**c**). The normalized asymmetry is the difference between the concentrations of dopants on either side of the well, divided by the total dopant concentration. The solid lines are fits to the theory of ref. 15 (see text). **b**, **d**, Plots summarizing the spin–orbit parameters from these fits in **a** and **c**. As in ref. 15, the spin–orbit coupling strengths are expressed as dimensionless quantities by normalizing to the Fermi velocity.

well, with both α and β_3 remaining essentially constant, indicating that we have varied β_1 independently. The spin–orbit parameters for each sample in the series are plotted in Fig. 2d. The crossing of α and β_1 occurs at a well width near 12 nm; however, the largest τ_E value occurs in the 11-nm sample. This finding is consistent with theory[15], which predicts that for constant D_S the peak lifetime occurs for $\alpha = \beta_1 - \beta_3$ rather than for $\alpha = \beta_1$. (When τ_E is normalized to account for variations in D_S from one sample to another, the same condition holds for asymmetry series as well; see Supplementary Information.)

As a check on the modelling of our TSG data, we compare the experimental values of α, β_1 and β_3 with band structure calculations. The Rashba coupling strength is predicted to obey $\alpha = r_{41}^{6c6c} e\langle E_z\rangle/\hbar v_F$, where e is the elementary charge, $\langle E_z\rangle$ is the average electric field in the well and r_{41}^{6c6c} is an intrinsic proportionality factor. In **k•p** perturbation theory, this factor is found to be 5.206 Å^2 for GaAs (ref. 19). To make a comparison with theory, we assume that the electrons in the well experience the delta layers as an infinite sheet of positive charge. We estimate the field strength to be 5.4×10^6 V m^{-1} for a normalized asymmetry of one. From the corresponding value of α, we find that $r_{41}^{6c6c} = 6.7$ Å^2, in good agreement with the perturbation theory result.

The Dresselhaus couplings, β_1 and β_3, are both proportional to a single intrinsic parameter, γ, with the linear term given by $\beta_1 = \gamma\langle k_z^2\rangle k_F/2E_F$, where E_F is the Fermi energy. From the values of β_1 as a function of well width, determined by TSG spectroscopy and analysis using the theory of ref. 15, we estimate that $\gamma = 5.0$ eV Å^3, assuming that $\langle k_z^2\rangle = (\pi/d)^2$. A larger value of γ would be obtained if $\langle k_z^2\rangle$ were reduced by penetration of the electron wavefunction into the barrier. Theoretical calculation of γ has proven to be challenging, and a wide range of values, 6.5–30 eV Å^3, have been reported for bulk GaAs (refs 20, 21). Comparison with experiment is further complicated by the existence of an interface Dresselhaus term, which, although often neglected, may be important in two-dimensional structures such as those studied here[22]. In the light of these complications, the value of γ that we obtain is in reasonable agreement with theory. It is also important to note that the experiments reported here potentially offer heightened sensitivity to the cubic Dresselhaus interaction, as proximity to the SU(2) point effectively eliminates spin relaxation from the linear terms. Independent of the value of γ, the ratio β_3/β_1 is given theoretically by $k_F^2/4\langle k_z^2\rangle$; that is, it is proportional to the ratio of the electron kinetic energies respectively parallel and perpendicular to the conducting plane. Again estimating that $k_z = \pi/d$, we find that the expected ratio for an 11-nm well is 0.16, consistent with our experimental value of 0.2. This agreement between theory and experiment supports the notion that the cubic Dresselhaus interaction limits the PSH lifetime in our samples at low temperature.

The T dependence of the spin-helix lifetimes further tests our understanding of 2DEG spin physics, and also is relevant to potential spintronics applications. The T dependence of the lifetime of each mode, for the sample closest to the SU(2) point, is plotted on a logarithmic scale in Fig. 3a. The lifetime of the spin–orbit-enhanced mode increases with decreasing T to ~50 K, then drops rapidly with further lowering of T. The lifetime of the spin–orbit-reduced mode, on the other hand, decreases monotonically with decreasing T.

We cannot rely solely on the spin dynamics theories described previously to explain the observed T dependence, as they consider only the $T = 0$ limit. However, these theories do indicate an important first step in the analysis. For a given set of spin–orbit parameters, the spin-helix lifetimes for both senses of rotation, and for all q, are predicted to scale as D_S^{-1}. Because D_S is known to depend strongly on T (ref. 23), this scaling provides at least one well-understood mechanism for T-dependent lifetimes. However, the fact that the enhanced- and reduced-lifetime modes display very different temperature dependences indicates immediately that scaling by $D_S(T)$ cannot fully account for the effect of temperature.

To focus on the T-dependent effects other than scaling by D_S, we consider the dimensionless parameter $\eta \equiv D_S q_{PSH}^2 \tau_{PSH}$, rather than

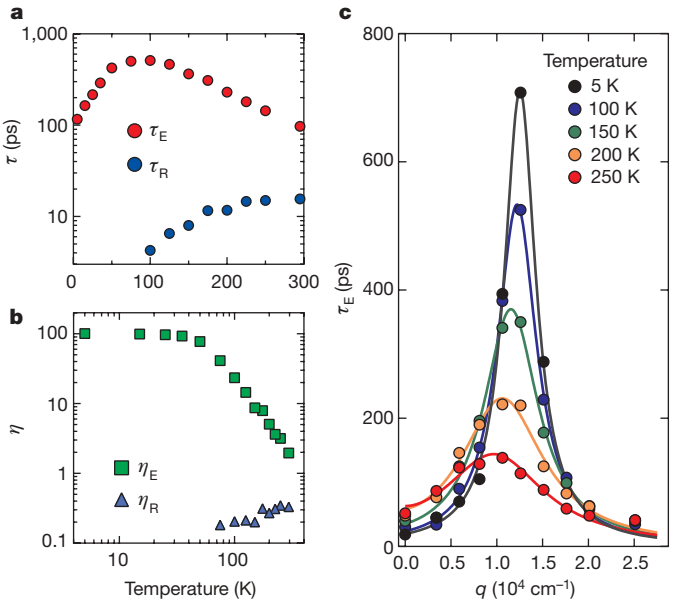


Figure 3 | Temperature dependence of the PSH. a, Temperature dependence of the lifetime of each helix mode at q_{PSH} for the 11-nm, asymmetrically doped sample, which is the closest to the SU(2) point. **b**, Temperature dependence of the dimensionless lifetime-enhancement factor $\eta \equiv D_S q_{PSH}^2 \tau_{PSH}$ for each helix mode. **c**, Temperature dependence of the PSH dispersion curves for a similar sample with a slightly reduced mobility. The reduced mobility suppresses the drop in D_S and τ by avoiding the ballistic crossover. Fits to the theory of ref. 15 (solid lines) are also shown.

τ_{PSH} itself (here the subscript PSH refers to quantities evaluated at the PSH wavevector). The parameter η is the measured PSH lifetime normalized to the lifetime predicted in the absence of spin–orbit coupling (see Methods), that is, a direct measure of the lifetime enhancement as a result of proximity to the SU(2) point. Figure 3b is a logarithmic plot of $\eta(T)$ for both helix modes. The enhancement factor corresponding to the rapidly decaying mode, η_R, is essentially independent of T, indicating that the temperature dependence of τ_R is entirely accounted for by $D_S(T)$. (The rapid increase of D_S below 75 K is the result of the quenching of the spin-Coulomb-drag effect[23,24].) By contrast, η_E is strongly T dependent. In this case, normalization with respect to $D_S(T)$ converts the peak in τ_E near 75 K to a plateau at low values of T, demonstrating that the PSH-lifetime enhancement is a monotonically decreasing function of increasing T. The enhancement factor is approximately 100 at low values of T and decays towards unity (no spin–orbit enhancement) at high temperature. Above 50 K, the enhancement factor obeys the power law $\eta(T) \propto T^{-2.2}$. In Fig. 3c, we plot the spin-lifetime dispersion curves for several values of T, illustrating the damping of the entire PSH resonance with increasing temperature. Although weakened, the PSH remains observable at room temperature. The existence of the PSH at temperatures far greater than that equivalent to the spin–orbit spin-splitting energy, which is only ~1 K, supports the idea that the lifetime enhancement is symmetry driven.

Why the PSH stability decreases strongly with T remains an open question. Within a non-interacting (or single-particle) picture, the cubic Dresselhaus term is the only SU(2)-breaking interaction. In the two-dimensional limit, where $\beta_3 \ll \beta_1$, the cubic Dresselhaus term can be viewed as a small, velocity-dependent correction to β_1, that is, $\beta_1^{eff} = \beta_1 - \beta_3 v^2/v_F^2$ (ref. 25). With increasing T, the thermal average of β_1^{eff} will decrease, driving the effective spin–orbit Hamiltonian farther from the SU(2) point. However, it is not clear at present whether this effect is sufficiently strong to account for the T dependence depicted in Fig. 3. A relatively weak reduction in PSH stability with increasing T was observed in numerical simulations of the spin kinetic equations for the specific case of $\alpha = 0.3\beta_1$ (ref. 25), and simulations with $\alpha \approx \beta_1$

have not yet been reported. To estimate the T dependence expected in this regime, we can substitute the thermal average of $\beta_1^{\rm eff}$ for β_1 in the formulae of ref. 15 for the PSH lifetime. Although we do obtain $\eta(T) \propto T^{-2}$ in the high-temperature limit, the onset of T^{-2} dependence is at approximately the Fermi temperature, which is ~400 K for our quantum wells. As the measured onset of the $T^{-2.2}$ dependence is at roughly 50 K, the nonlinear velocity dependence of $\beta_1^{\rm eff}$ may not fully account for the reduction in PSH lifetime with temperature. In considering effects beyond the single-particle picture, the approximate T^{-2} scaling suggests a connection with electron–electron scattering. As mentioned previously, if SU(2) symmetry is exact then the PSH lifetime is not sensitive to electron–electron scattering[2]. However, it remains to be seen whether many-body interactions can affect the PSH lifetime when SU(2) is weakly broken by the cubic Dresselhaus term, disorder in local spin–orbit couplings[26] or spin-dependent scattering mechanisms.

Finally, we note that a PSH-lifetime enhancement of 100 is not a fundamental limit. When controlled by the cubic Dresselhaus term, the lifetime enhancement is proportional to $(\beta_1/\beta_3)^2$. Gated structures in which electron density and electric field are tuned independently will enable this ratio to be increased, while maintaining $\alpha = \beta_1$. Increased stability of the PSH creates possibilities for new experiments on spin transport, such as the measurement of the intrinsic spin Hall effect, the study of charge transport dynamics in the presence of strong spatial variation of spin polarization and the demonstration of efficient spin transistors.

METHODS SUMMARY

The GaAs/$Al_{0.3}Ga_{0.7}As$ quantum-well samples were grown on semi-insulating GaAs in the [001] direction by molecular beam epitaxy, and consisted of ten quantum wells separated by 48-nm barriers. The Si donors were deposited in eight single-atomic layers in the central 14 nm of each barrier to maximize their distance from the 2DEG. To tune α, the ratio of the donor concentrations in alternating barriers was adjusted by varying the Si deposition times, and the total target carrier concentration in the wells was held fixed at $n = 8 \times 10^{11}\,{\rm cm}^{-2}$. The electron mobility typically reached $\mu \approx 3 \times 10^5\,{\rm cm}^2\,{\rm V}^{-1}\,{\rm s}^{-1}$ at low temperature. All samples were mounted on *c*-axis-cut sapphire discs, and the GaAs substrates were chemically etched to allow for spin-grating measurements in transmission geometry.

We determined the spin diffusion coefficient, $D_{\rm S}$, through analysis of the uniform spin polarization, $S_z(q = 0, t)$, measured by a standard time-resolved Faraday rotation technique[10]. As T was reduced from room temperature, $S_z(q = 0, t)$ crossed over from single-exponential decay to damped oscillations (at about 50 K in our quantum wells). The crossover occurred when the electron mean free time became comparable to the period of precession in the spin–orbit effective fields. In the high-T regime, we determined $D_{\rm S}$ from $1/\tau_{\rm S} = D_{\rm S} q_{\rm PSH}^2$ (ref. 27), which holds in the regime where α and β_1 are approximately equal. Here $\tau_{\rm S}$ is the $q = 0$ spin lifetime. To determine $D_{\rm S}$ through the crossover regime, we used the phenomenological formula

$$S_z(q=0,\,t) \propto \int_0^{\pi/2} \exp\left[-\left(\frac{\Omega\tau\cos\phi}{1-i\Omega\tau\cos\phi}\right)\Omega t\right]\,{\rm d}\phi$$

where τ is the mean free time and $\Omega\cos\phi$ is the precession frequency as function of angle, ϕ, on the Fermi circle. This expression interpolates between the exact result in the $\Omega\tau \gg 1$ limit, which is the zero-order Bessel function, and the non-oscillatory decay in the $\Omega\tau \to 0$ limit. We verified that the values of $D_{\rm S}$ obtained from the $q = 0$ data are consistent with those obtained from analysis of the full q dependence of the spin lifetimes.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 October 2008; accepted 30 January 2009.

1. Noether, E. Invariante Variationsprobleme. *Nachr. König. Gesellsch. Wiss. Göttingen, Math-Phys. Klasse* 235–257 (1918).
2. Bernevig, B. A., Orenstein, J. & Zhang, S.-C. Exact SU(2) symmetry and persistent spin helix in a spin-orbit coupled system. *Phys. Rev. Lett.* **97,** 236601 (2006).
3. Bychkov, Y. A. & Rashba, E. I. Oscillatory effects and the magnetic susceptibility of carriers in inversion layers. *J. Phys. Chem.* **17,** 6039–6045 (1984).
4. Dresselhaus, G. Spin–orbit coupling effects in zinc blende structures. *Phys. Rev.* **100,** 580–586 (1955).
5. Cameron, A. R., Riblet, P. & Miller, A. Spin gratings and the measurement of electron drift mobility in multiple quantum well semiconductors. *Phys. Rev. Lett.* **76,** 4793–4796 (1996).
6. Awschalom, D. D., Loss, D. & Samarth, N. (eds) *Semiconductor Spintronics and Quantum Computation* (Springer, 2002).
7. Ohno, M. & Yoh, K. Datta-Das-type spin-field-effect transistor in the nonballistic regime. *Phys. Rev. B* **77,** 045323 (2008).
8. Meier, F. & Zakharchenya, B. *Optical Orientation* (North-Holland, 1984).
9. Gedik, N. & Orenstein, J. Absolute phase measurement in heterodyne detection of transient gratings. *Opt. Lett.* **29,** 2109–2111 (2004).
10. Crooker, S. A., Awschalom, D. D. & Samarth, N. Time-resolved Faraday rotation spectroscopy of spin dynamics in digital magnetic heterostructures. *IEEE J. Sel. Top. Quantum Electron.* **1,** 1082–1092 (1995).
11. Froltsov, V. A. Diffusion of inhomogeneous spin distribution in a magnetic field parallel to interfaces of a III–V semiconductor quantum well. *Phys. Rev. B* **64,** 045311 (2001).
12. Burkov, A. A., Nunez, A. S. & MacDonald, A. H. Theory of spin-charge-coupled transport in a two-dimensional electron gas with Rashba spin-orbit interactions. *Phys. Rev. B* **70,** 155308 (2004).
13. Weber, C. P. *et al.* Nondiffusive spin dynamics in a two-dimensional electron gas. *Phys. Rev. Lett.* **98,** 076604 (2007).
14. Schliemann, J., Egues, J. C. & Loss, D. Nonballistic spin-field-effect transistor. *Phys. Rev. Lett.* **90,** 146801 (2003).
15. Stanescu, T. D. & Galitski, V. Spin relaxation in a generic two-dimensional spin-orbit coupled system. *Phys. Rev. B* **75,** 125307 (2007).
16. Braun, W. Trampert, A. Däweritz, L. & Ploog, K. H. Nonuniform segregation of Ga at AlAs/GaAs heterointerfaces. *Phys. Rev. B* **55,** 1689–1695 (1997).
17. de Andrada e Silva, E. A. La Rocca, G. C. & Bassani, F. Spin-orbit splitting of electronic states in semiconductor asymmetric quantum wells. *Phys. Rev. B* **55,** 16293–16299 (1997).
18. Schubert, E. F. *et al.* Fermi-level-pinning-induced impurity redistribution in semiconductors during epitaxial growth. *Phys. Rev. B* **42,** 1364–1368 (1990).
19. Winkler, R. *Spin–Orbit Coupling Effects in Two-Dimensional Electron and Hole Systems* (Springer Tracts Mod. Phys. Vol. 191, Springer, 2003).
20. Krich, J. J. & Halperin, B. I. Cubic Dresselhaus spin-orbit coupling in 2D electron quantum dots. *Phys. Rev. Lett.* **98,** 226802 (2007).
21. Chantis, A. N., Schilfgaarde, M. & Kotani, T. *Ab initio* prediction of conduction band spin splitting in zinc blende semiconductors. *Phys. Rev. Lett.* **96,** 086405 (2006).
22. Fabian, J., Matos-Abiague, A., Ertler, C., Stano, P. & Zutic, I. Semiconductor spintronics. *Acta Physica Slovaca* **57,** 565–907 (2007).
23. D'Amico, I. & Vignale, G. Spin Coulomb drag in the two-dimensional electron liquid. *Phys. Rev. B* **68,** 045307 (2003).
24. Weber, C. P. *et al.* Observation of spin Coulomb drag in a two-dimensional electron gas. *Nature* **437,** 1330–1333 (2005).
25. Weng, M. Q., Wu, M. W. & Cui, H. L. Spin relaxation in *n*-type GaAs quantum wells with transient spin grating. *J. Appl. Phys.* **103,** 063714 (2008).
26. Sherman, E. & Ya.. Random spin–orbit coupling and spin relaxation in symmetric quantum wells. *Appl. Phys. Lett.* **82,** 209–211 (2003).
27. D'Yakonov, M. I., &. Perel', V. I. Spin relaxation of conduction electrons in noncentrosymmetric semiconductors. *Sov. Phys. Solid State* **13,** 3023–3026 (1971).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Work performed at Lawrence Berkeley National Laboratory and Stanford University was supported by the US Department of Energy, Office of Basic Energy Science, Materials Science and Engineering Division, and at the University of California, Santa Barbara by the US National Science Foundation and Office of Naval Research. S.M. acknowledges partial support through the National Defense Science and Engineering Graduate Fellowship Program. We thank J. Stephens and J. Krich for discussions, G. Fleming for use of a phase-mask array, and K. Bruns for creating the PSH diagram of Fig. 1c.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.D.K. (jdkoralek@lbl.gov).

METHODS

Transient spin gratings. Transient spin polarization waves were generated by the optical interference of two cross-polarized pulses from a single mode-locked Ti:sapphire laser (80 MHz, 100 fs), focused non-collinearly onto the 2DEG. The pump pulses were amplitude-modulated at 100 kHz using a photo-elastic modulator. The time evolution of the spin polarization was monitored by time-delayed probe pulses, which see the modulation of the 2DEG polarization as a diffraction grating because of the Kerr effect. The amplitude and phase of the transient spin grating were measured using a heterodyne detection scheme. The diffracted pulses were mixed at a Si photodiode detector with another beam from the same laser, which served as a local oscillator[9]. The relative phase of the signal and local oscillator pulses was modulated at 210 Hz by transmitting one beam through a coverslip mounted on a torsional oscillator. Synchronous detection of the mixed signal was accomplished using two lock-in amplifiers, the first referenced to the pump amplitude-modulation frequency and the second referenced to the modulation frequency of the relative phase.

LETTERS

Solubility trapping in formation water as dominant CO_2 sink in natural gas fields

Stuart M. V. Gilfillan[1,2], Barbara Sherwood Lollar[3], Greg Holland[1], Dave Blagburn[1], Scott Stevens[4], Martin Schoell[5], Martin Cassidy[6], Zhenju Ding[1,7], Zheng Zhou[1], Georges Lacrampe-Couloume[3] & Chris J. Ballentine[1]

Injecting CO_2 into deep geological strata is proposed as a safe and economically favourable means of storing CO_2 captured from industrial point sources[1–3]. It is difficult, however, to assess the long-term consequences of CO_2 flooding in the subsurface from decadal observations of existing disposal sites[1,2]. Both the site design and long-term safety modelling critically depend on how and where CO_2 will be stored in the site over its lifetime[2–4]. Within a geological storage site, the injected CO_2 can dissolve in solution or precipitate as carbonate minerals. Here we identify and quantify the principal mechanism of CO_2 fluid phase removal in nine natural gas fields in North America, China and Europe, using noble gas and carbon isotope tracers. The natural gas fields investigated in our study are dominated by a CO_2 phase and provide a natural analogue for assessing the geological storage of anthropogenic CO_2 over millennial timescales[1,2,5,6]. We find that in seven gas fields with siliciclastic or carbonate-dominated reservoir lithologies, dissolution in formation water at a pH of 5–5.8 is the sole major sink for CO_2. In two fields with siliciclastic reservoir lithologies, some CO_2 loss through precipitation as carbonate minerals cannot be ruled out, but can account for a maximum of 18 per cent of the loss of emplaced CO_2. In view of our findings that geological mineral fixation is a minor CO_2 trapping mechanism in natural gas fields, we suggest that long-term anthropogenic CO_2 storage models in similar geological systems should focus on the potential mobility of CO_2 dissolved in water.

Noble gas and CO_2 carbon isotopes are powerful tracers of crustal fluid processes that act on subsurface CO_2 (refs 5, 7–10). Within a geological storage site, CO_2 injected as a free CO_2 phase (gas or supercritical) may over time be dissolved in solution (solubility trapping), or locked within carbonate minerals by precipitation (mineral trapping)[4,11]. By using noble gas and carbon isotope tracers together to study naturally occurring CO_2 systems, we can uniquely identify and quantify the principal mechanism of the CO_2 phase removal (mineral or solubility trapping) over a timescale not accessible through extant injection studies.

We combine noble gas data from five natural CO_2 reservoirs located within the Colorado Plateau and Rocky Mountain provinces (McCallum dome, Sheep Mountain and McElmo dome, in Colorado; Bravo dome, in New Mexico; and St Johns dome, in Arizona and New Mexico)[7] with new $\delta^{13}C(CO_2)$ isotope data (Table 1). Previous work has shown that noble gas patterns in these gas fields are explained by the stripping of CO_2 gas from the formation water during reservoir filling, followed by partial dissolution of noble gases back into the formation water[7]. We also consider published noble gas and stable isotope information in a further four CO_2-rich natural gas fields (the JM-Brown Bassett (JMBB) field in the Permian basin, Texas[5]; the Kismarja field in the Pannonian basin, Hungary[8]; and the Jilin field in the Songliao basin, Jilin Province, and the Subei basin field, Jiangsu Province, in China[12,13]).

$CO_2/^3He$ ratios within the magmatic range of $(1–10) \times 10^9$ have been used to identify a primary magmatic origin of the CO_2 contained within five natural CO_2 reservoirs of the Colorado Plateau and Rocky Mountain provinces[7]. $CO_2/^3He$ ratios within the Subei basin and the JMBB field also indicate a magmatic origin, but the $CO_2/^3He$ ratios within the Jilin and Kismarja fields are much higher, suggesting a predominantly crustal origin[5,8,12,13]. All of the reservoirs exhibit local variation in the CO_2 content relative to the inert tracer ^3He. As there is not a significant source of ^3He within the crust[14], and as ^3He is inert and highly insoluble[9], this variation must be due to changes in the CO_2 component within the reservoirs. Although many sources and sinks of CO_2 exist in the subsurface[4,8,9], below we argue that the variation in $CO_2/^3He$ ratios is caused by CO_2 loss from the reservoir. The difference between the highest $CO_2/^3He$ ratio and lower values can provide a minimum estimate of this CO_2 loss. In the case of Bravo dome, a reduction in $CO_2/^3He$ values from 4.82×10^9 (well BD11) to 2.25×10^9 (BD02) indicates a loss of the original CO_2 charge of >50% in the portion of the reservoir represented by BD02 (Table 1). Samples from McElmo dome show a decrease from 8.5×10^9 (YD-1) to 0.68×10^9 (HE-2), suggesting a loss of emplaced CO_2 of >90% in portions of this field.

^4He is continually produced in the subsurface by the radiogenic decay of U, Th and K (ref. 14). ^{20}Ne is introduced into the subsurface as a component of air dissolved in water and, as such, can only enter the reservoir system through interaction with formation water[9]. Although there is no *a priori* reason to expect a correlation between ^4He and ^{20}Ne, one has been observed in natural gases on a regional scale[15]. This correlation is the result of ^4He accumulating in the formation water[16], which also contains atmosphere-derived ^{20}Ne, and subsequent quantitative partitioning of both ^4He and ^{20}Ne into the reservoir phase[7,15]. Almost all CO_2 reservoirs for which we have ^{20}Ne and ^4He concentration data show a local ^{20}Ne correlation with ^4He (Table 1 and Supplementary Information). A decrease in $CO_2/^3He$ is also correlated with ^{20}Ne in most CO_2 reservoirs (Fig. 1) and with ^4He in all CO_2 reservoirs (Fig. 2).

There are various mechanisms by which crustal CO_2 ($CO_2/^3He > 10^{10}$) can be added to these systems[4,10], but there is no plausible mechanism that enables crustal CO_2 to be variably added to these systems while preserving a correlation of $CO_2/^3He$ with the noble gases derived from formation water. Neglecting small amounts of ^3He dissolution back into the formation water[7], changes in

[1]School of Earth, Atmospheric and Environmental Sciences, The University of Manchester, Oxford Road, Manchester M13 9PL, UK. [2]Scottish Centre for Carbon Storage, School of GeoSciences, The University of Edinburgh, Grant Institute, Kings Buildings, West Mains Road, Edinburgh EH9 3JW, UK. [3]Department of Geology, University of Toronto, 22 Russell Street, Toronto, Ontario M5S 3B1, Canada. [4]Advanced Resources International, 4501 Fairfax Drive, Suite 910, Arlington, Virginia 22203-1661, USA. [5]GasConsult International, 2808 Adeline Street #3, Berkeley, California 94703, USA. [6]Department of Earth and Atmospheric Sciences, University of Houston, Houston, Texas 77204-5503, USA. [7]China University of Geosciences, Wuhan City, 430074, China.

Table 1 | Sample location, producing formation, major gas species and CO_2 carbon isotopes

Field and well	Location	Producing formation	$CO_2/^3He$ (10^9)	$^3He/^4He$ (R/R_a)	^4He (10^{-4} cm^3 (STP) cm^{-3})	^{20}Ne (10^{-8} cm^3 (STP) cm^{-3})	$\delta^{13}C(CO_2)$ (‰)
Bravo dome[7]							
BD01	35.8613, −103.2947	Tubb	4.53 (10)	1.670 (8)	0.944 (12)	0.169 (2)	−3.96 (4)
BD02	36.0058, −103.2305	Tubb	2.25 (5)	0.764 (4)	4.15 (5)	0.700 (7)	−4.93 (8)
BD03	36.0934, −103.2662	Tubb	2.41 (5)	0.896 (4)	3.31 (4)	0.521 (5)	−4.89 (19)
BD04	35.9766, −103.3480	Tubb	4.61 (10)	1.611 (8)	0.961 (2)	0.181 (2)	−4.23 (8)
BD05	35.9190, −103.2059	Tubb	2.74 (6)	0.965 (5)	2.70 (4)	0.446 (4)	−4.95 (5)
BD06	36.1080, −103.4988	Tubb	3.94 (8)	1.503 (8)	1.20 (2)	0.202 (2)	−4.55 (11)
BD07	35.9046, −103.4190	Tubb	4.34 (9)	2.104 (11)	0.781 (10)	0.180 (2)	−4.85 (1)
BD08	35.8037, −103.4368	Tubb	3.87 (8)	1.143 (6)	1.61 (2)	0.264 (3)	−3.88 (8)
BD09	36.0496, −103.4452	Tubb	4.22 (9)	1.724 (9)	0.981 (12)	0.180 (2)	−4.44 (11)
BD10	36.1519, −103.3557	Tubb	3.25 (6)	1.104 (6)	1.99 (3)	0.308 (3)	−4.88 (7)
BD11	35.8469, −103.7032	Tubb	4.82 (10)	3.784 (19)	0.391 (5)	0.103 (1)	−3.66 (29)
BD12	35.8469, −103.7387	Tubb	4.74 (10)	3.627 (18)	0.415 (6)	NM	−3.94 (17)
BD13	35.7749, −103.2059	Tubb	3.54 (8)	1.318 (7)	1.53 (2)	0.240 (3)	−4.42 (3)
BD14	35.7893, −103.3302	Tubb	4.39 (9)	1.413 (7)	1.15 (2)	0.179 (4)	−4.04 (2)
BD12b	35.8469, −103.7387	Tubb	4.75 (10)	3.634 (18)	0.413 (6)	0.120 (2)	−3.94 (17)
McCallum dome[7]							
No. 3 (8-3)	40.7632, −106.1717	Lakota	1.52 (4)	0.354 (7)	12.3 (2)	1.17 (2)	−5.1 (3)
No. 5	40.7777, −106.2479	Lakota	1.04 (3)	0.409 (7)	15.5 (2)	2.71 (3)	−5.2 (1)
No. 13	40.7777, −106.2289	Lakota/Morrison	0.89 (2)	0.393 (7)	18.8 (2)	4.36 (5)	−5.3 (2)
No. 79	40.7777, −106.2670	Dakota/Lakota	1.77 (6)	0.406 (6)	9.16 (21)	2.53 (3)	−5.7 (1)
McElmo dome[7]							
MC-1	37.4155, −108.7713	Leadville	5.04 (11)	0.145 (2)	9.58 (8)	0.376 (4)	−4.26 (10)
HE-2	37.5052, −108.9094	Leadville	0.68 (15)	0.148 (1)	70.5 (7)	0.307 (30)	−4.40 (10)
YC-4	37.4529, −108.8583	Leadville	4.96 (11)	0.137 (3)	10.2 (10)	0.573 (6)	−4.41 (10)
SC-9	37.3934, −108.8733	Leadville	3.17 (7)	0.150 (3)	14.8 (14)	0.497 (5)	−4.29 (10)
YB-2	37.4472, −108.8075	Leadville	8.74 (20)	0.125 (1)	6.42 (61)	0.371 (4)	−4.40 (10)
YC-1	37.4529, −108.8583	Leadville	4.07 (9)	0.142 (2)	12.1 (12)	0.423 (5)	−4.34 (10)
HF-1	37.4871, −108.8807	Leadville	2.16 (6)	0.169 (1)	19.3 (26)	0.564 (12)	−4.37 (10)
HD-2	37.4572, −108.9008	Leadville	4.28 (10)	0.140 (3)	11.7 (12)	0.128 (2)	−4.38 (10)
YA-2	37.4692, −108.7811	Leadville	3.39 (8)	0.138 (3)	15.0 (15)	0.130 (2)	−4.42 (10)
YE-1	37.4818, −108.8123	Leadville	4.16 (9)	0.173 (3)	9.75 (8)	0.143 (3)	−4.45 (10)
HA-1	37.5289, −108.8718	Leadville	4.56 (10)	0.139 (3)	11.0 (11)	0.205 (7)	−4.66 (10)
SC-10	37.3934, −108.8733	Leadville	4.37 (10)	0.139 (2)	11.6 (11)	0.413 (5)	−4.27 (10)
HC-2	37.4734, −108.8860	Leadville	4.68 (11)	0.140 (2)	10.7 (10)	0.409 (5)	−4.38 (10)
HB-1	37.5087, −108.8802	Leadville	4.74 (11)	0.148 (3)	9.94 (10)	0.247 (4)	−4.49 (10)
YD-1	37.4619, −108.8224	Leadville	8.50 (20)	0.145 (3)	5.68 (6)	0.366 (5)	−4.46 (10)
JM-Brown Basset field[5]							
Turk State No. 1A	30.38758, −101.85642	Ellenberger	5.92 (47)	0.543 (16)	1.25 (9)	NM	−2.88 (3)
Bassett Goode No. 3	30.37852, −101.83068	Ellenberger	5.55 (43)	0.527 (16)	1.42 (10)	NM	−2.89 (3)
Brown Bassett No. 2*	30.34433, −101.7995	Ellenberger	5.82 (35)	0.502 (15)	1.33 (7)	NM	−2.90 (3)
Mayme K. Martin ETAL 1	30.35661, −101.74721	Ellenberger	5.29 (40)	0.372 (11)	1.42 (10)	NM	−2.97 (3)
Mitchell 109 No. 2*	30.33329, −101.69826	Ellenberger	4.58 (36)	0.400 (12)	1.53 (11)	NM	−2.92 (3)
Mitchell 5 No. 1X	30.32352, −101.68429	Ellenberger	5.61 (43)	0.478 (11)	1.40(10)	NM	−2.84 (3)
Mitchell 103 No. 2	30.3568, −101.63642	Ellenberger	4.20 (33)	0.246 (7)	1.39 (10)	NM	−2.70 (3)
Mitchell No. 6	30.351, −101.58835	Ellenberger	3.93 (31)	0.264 (8)	1.51 (11)	NM	−2.96 (3)
Mitchell No. 3	30.33966, −101.61307	Ellenberger	4.22 (33)	0.240 (7)	1.39 (10)	NM	−3.06 (3)
Mitchell A-11 No. 1	30.30286, −101.57677	Ellenberger	4.07 (32)	0.272 (8)	1.66 (12)	NM	−2.93 (3)
Mitchell No. 12	30.29118, −101.57295	Ellenberger	4.24 (130)	0.267 (8)	1.46 (10)	NM	−2.96 (3)
Sheep Mountain[7]							
8-2-P	37.6383, −105.1836	Dakota	2.31 (5)	0.981 (10)	3.13 (3)	1.47 (2)	−5.0 (2)
2-10-O	37.6966, −105.2018	Entrada	2.44 (6)	0.984 (12)	2.96 (3)	3.04 (3)	−5.2 (1)
9-26	37.6675, −105.1836	Dakota	2.57 (6)	0.934 (14)	2.95 (3)	0.613 (9)	NM
2-9-H	37.7112, −105.2200	Dakota	2.44 (6)	0.945 (19)	3.07 (3)	9.77 (10)	NM
3-15-B	37.6966, −105.2018	Dakota	2.61 (6)	0.937 (16)	2.90 (3)	1.54 (2)	−5.7 (4)
4-13	—	Dakota	2.17 (5)	0.942 (18)	3.47 (4)	1.11 (2)	NM
4-26-E	37.6675, −105.1836	Entrada	2.20 (5)	1.024 (18)	3.15 (3)	0.442 (4)	−4.8 (1)
3-23-D	37.6820, −105.2018	Dakota	2.26 (5)	0.988 (14)	3.17 (3)	0.579 (9)	NM
7-35-L	37.6383, −105.1836	Dakota	2.53 (6)	0.916 (14)	3.06 (3)	0.749 (12)	−5.0 (2)
2-35-C	37.6675, −105.1836	Dakota	2.57 (6)	0.963 (19)	2.87 (3)	0.573 (8)	NM
1-15-C	37.6966, −105.2018	Entrada	2.71 (6)	0.967 (16)	2.71 (3)	6.77 (10)	NM
3-4-O	37.7112, −105.2200	Dakota	2.53 (6)	0.937 (14)	2.99 (3)	2.64 (3)	−5.8 (3)
4-14-M	37.6820, −105.2018	Dakota	2.65 (6)	0.892 (15)	3.00 (3)	1.11 (1)	NM
5-15-O	37.6820, −105.2018	Dakota	2.30 (5)	1.056 (15)	2.92 (3)	4.33 (5)	−5.0 (1)
4-4-P	37.7112, −105.2200	Dakota	2.90 (7)	0.970 (14)	2.52 (2)	1.31 (2)	NM
5-9-A	37.7112, −105.2200	Dakota	2.39 (6)	1.006 (18)	2.94 (3)	1.28 (2)	NM
1-1-J	37.6383, −105.1836	Dakota	3.61 (8)	0.908 (16)	2.16 (2)	0.878 (12)	−5.2 (1)
1-22-H	37.5946, −105.2018	Entrada	2.25 (5)	0.981 (17)	3.22 (3)	0.937 (13)	−4.5 (2)
St Johns dome[7]							
22-1X	34.4265, −109.2664	Supai	0.098 (2)	0.455 (8)	134 (13)	34.4 (47)	−3.65 (5)
10-22	34.2437, −109.1645	Supai	1.91 (42)	0.394 (8)	9.42 (9)	2.30 (4)	−3.79 (5)
3-1	34.3771, −109.2563	Supai	0.22 (3)	0.433 (9)	70.6 (7)	15.1 (21)	−3.85 (5)
Jilin field[12,13,21]							
Wan 2	—	Cretaceous	1.44 (4)	4.91 (6)	1.00 (2)	NM	−3.6
Wan 5	—	Cretaceous	227 (7)	4.10 (4)	0.0076 (2)	0.0547 (15)	−5.0
Wan 6	—	Cretaceous	8.32 (3)	4.99 (5)	0.169 (4)	0.230 (6)	−3.8
Wan 8	—	Cretaceous	NM	4.30 (5)	NM	NM	−3.2

Table 1 is continued on page 616.

Table 1 | Continued

Field and well	Location	Producing formation	$CO_2/^3He$ (10^9)	$^3He/^4He$ (R/R_a)	^4He (10^{-4} cm^3 (STP) cm^{-3})	^{20}Ne (10^{-8} cm^3 (STP) cm^{-3})	$\delta^{13}C(CO_2)$ (‰)
Wan 9	—	Cretaceous	36.6 (10)	4.08 (4)	0.047 (1)	0.130 (3)	−3.8
Subei field[12,21]							
Huangqian 1	—	Permian	2.17 (7)	3.52 (5)	3.13 (3)	1.47 (2)	−3.6
Sutail 74	—	Devonian	0.493(14)	3.59 (4)	2.96 (3)	3.04 (3)	−4.1
Su203	—	Eocene	0.459 (13)	2.61 (3)	2.95 (3)	0.613 (9)	−2.7
Kismarja field[8,29]							
Kismarja 8	—	Up. Pannonian	20.2 (5)	1.33 (3)	0.226 (7)	NM	−5.0
Kismarja 79	—	Up. Pannonian	15.5 (4)	1.38 (3)	0.310 (10)	NM	−4.9
Kismarja 61	—	Up. Pannonian	27.3 (6)	1.16 (2)	0.205 (6)	NM	−5.1
Kismarja 55	—	Up. Pannonian	13.3 (3)	1.38 (3)	0.360 (11)	NM	−5.1
Kismarja 56	—	Up. Pannonian	1090 (3)	1.16 (2)	0.0052 (2)	NM	−6.8
Kismarja 74	—	Up. Pannonian	65.2 (2)	1.34 (3)	0.078 (3)	NM	−6.4
Kismarja 22	—	Up. Pannonian	1.52 (1)	1.02 (2)	1.31 (3)	NM	−6.6

Location is given as latitude and longitude in decimal degress, where north and east are positive and south and west are negative. The $^3He/^4He$ ratio, R, is shown relative to the $^3He/^4He$ ratio in air, R_a, which is taken to be 1.399×10^{-6}. $\delta^{13}C‰ = [(^{13}C/^{12}C_{sample} - {}^{13}C/^{12}C_{standard})/(^{13}C/^{12}C_{standard})] \times 1{,}000$; the standard used is the Vienna PeeDee Belemnite. Errors (1σ) are shown in parentheses. NM, not measured.
* Average of two analyses for He and Ar.

$CO_2/^3He$ must therefore be due to CO_2 loss in the subsurface by an amount directly proportional to the amount of formation water that has been degassed. CO_2 is soluble and reactive. The most probable mechanisms of subsurface CO_2 fluid phase removal are solubility and/or mineral trapping[4,11].

Reservoir lithology may exert a significant influence on how changes in $CO_2/^3He$ ratio relate to $\delta^{13}C(CO_2)$. The carbonate reservoirs (the JMBB field and the McElmo and St Johns domes) show little variance in $\delta^{13}C(CO_2)$, whereas the siliciclastic fields (the Jilin, Subei basin and Kismarja fields, Sheep Mountain, McCallum dome and Bravo dome) exhibit a greater $\delta^{13}C(CO_2)$ range (Table 1 and Supplementary Fig. 1). We consider Bravo dome and McElmo dome as representative cases for each type of reservoir lithology.

Emplacement of CO_2 at Bravo dome is believed to have occurred relatively recently (local volcanic activity dates from 8,000 to 10,000 years ago)[7,17], and the field may still be undergoing active CO_2 recharge[11]. The decreasing $CO_2/^3He$ ratio within Bravo dome correlates with more negative $\delta^{13}C(CO_2)$ (Fig. 3a). Taking the highest $CO_2/^3He$ ratio, of 4.82×10^9 (BD11), to be the sample that experienced

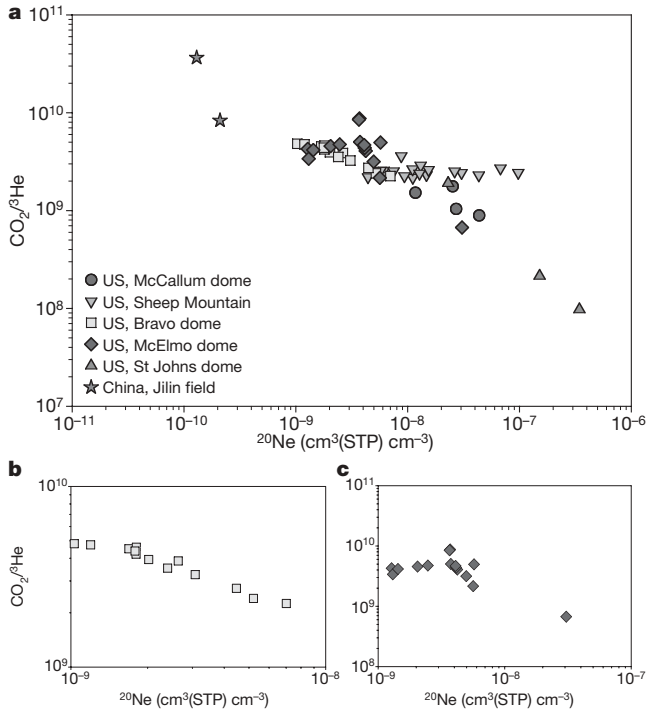


Figure 1 | $CO_2/^3He$ variation plotted against ^{20}Ne from CO_2-rich natural gas fields. **a**, There is a general trend in this data set of decreasing $CO_2/^3He$ with increasing ^{20}Ne. **b**, **c**, This trend is most clear in the siliciclastic-case data set, from Bravo dome (**b**), but less clear in the data from the carbonate-case reservoir, McElmo dome (**c**). ^3He is conservative within the gas phase. Lower $CO_2/^3He$ ratios therefore represent subsurface reduction in CO_2 concentration in the emplaced CO_2 phase. Because the only subsurface source of the ^{20}Ne is the formation water, the CO_2 sink must be linked to the formation water contacted by the gas phase. STP indicates measurement at the International Union of Pure and Applied Chemistry (IUPAC) standard temperature (0 °C) and pressure (100 kPa).

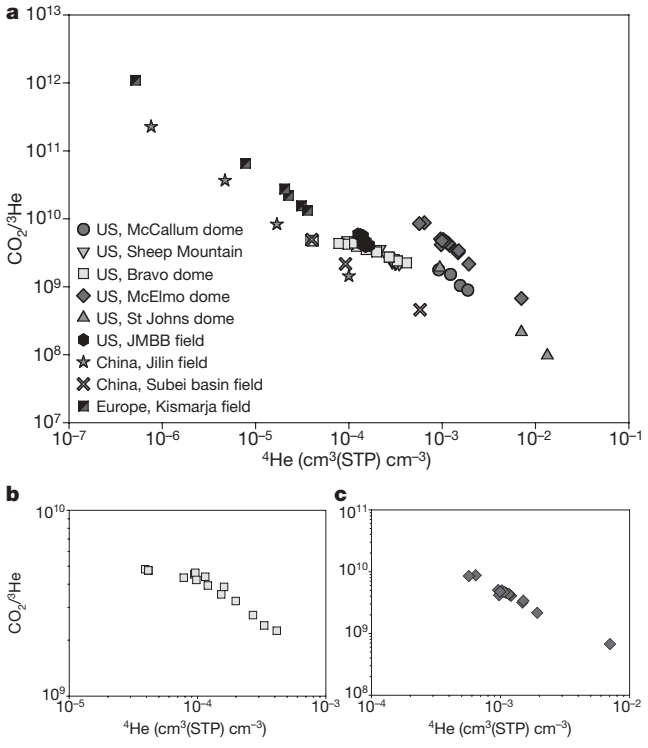


Figure 2 | $CO_2/^3He$ in CO_2-rich natural gas fields shows strong anticorrelation with ^4He. **a**, ^4He accumulates in formation water over time[7,15,16] and underscores the importance of formation water in controlling the mechanism of subsurface CO_2 removal (Fig. 1 and main text). We speculate that the formation-water ^4He signature with $CO_2/^3He$ is more coherent than the equivalent ^{20}Ne signature (Fig. 1) owing to perturbation of ^{20}Ne in ancient formation water through non-water phase interaction[9], with subsequent ^4He accumulation providing a homogenous, regional-scale formation-water ^4He signal[15,16]. Different $CO_2/^3He$-versus-^4He gradients are due to different local formation-water ^4He accumulation rates. **b**, **c**, As in Fig. 1, but for ^4He.

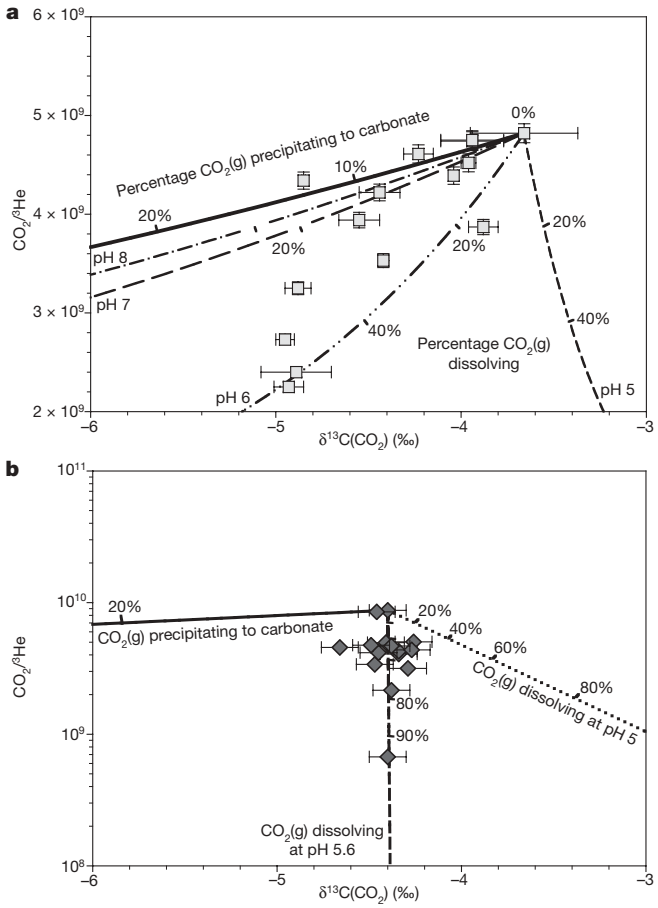


Figure 3 | Plot of $\delta^{13}C(CO_2)$ against $CO_2/^3He$ for Bravo dome and McElmo dome. **a**, Bravo Dome. The solid line shows the predicted trend for carbonate mineral precipitation and the broken lines show CO_2(g) dissolution trends for the indicated formation-water pH (see Methods Summary). This data limits the maximum effect of CO_2 precipitation in samples to approximately 18%. **b**, McElmo Dome. Invariant $\delta^{13}C(CO_2)$ with a change in $CO_2/^3He$ of over an order of magnitude in McElmo dome gases cannot be accounted for by precipitation (solid line). Dissolution of reservoir CO_2 into formation water at pH 5.6 is consistent with observed results. Error bars are 1σ.

the least CO_2 loss, we calculate the coherent change in $CO_2/^3He$ and $\delta^{13}C(CO_2)$ predicted for CO_2 dissolution into the formation water at various pH values and for CO_2 precipitation as a carbonate (Methods Summary). The data are not consistent with precipitation as carbonate being a major sink for CO_2 at Bravo dome (Fig. 3a). However, although a significant number of the data points are consistent with CO_2 dissolution into formation water at a pH between 6 and 7, it is not possible to rule out a degree of CO_2 loss due to precipitation together with CO_2 dissolution at a lower pH (for example pH 5). In such a two-process model, an upper limit of approximately 18% can be set on the proportion of CO_2 lost to precipitation (Fig. 3a). Hence, in all cases the major CO_2 sink is dissolution.

In situ precipitation of 18% of reservoir CO_2 would generate between 3.2 and 6.1% by mass of the whole rock, depending on whether dolomite, calcite or dawsonite precipitation was favoured by the reservoir conditions. Although evidence for CO_2-rich formation water interaction within the reservoir has been documented, so far no secondary carbonate has been identified[18]. Nevertheless, the volume control of the water suggests that the location of the precipitate, if any, is likely to be within the water leg that was not sampled. Lack of reservoir secondary mineralization cannot at this stage rule out any carbonate precipitation as a minor CO_2 sink.

Similar to the case for Bravo dome, although many of the Sheep Mountain data can be accounted for by dissolution of CO_2 (at pH 5 in this case), a small component of precipitation cannot be ruled out. Adopting the same approach as used for Bravo dome, we find that the remaining Sheep Mountain data require a maximum of 10% precipitation and 20% dissolution of the original CO_2 charge (Table 1 and Supplementary Fig. 2). By contrast, although minor data scatter may also be due to some small amount of CO_2 precipitation or dissolution at pH 7–8, almost all the data from the other siliciclastic fields (McCallum dome and the Subei basin, Kismarja and Jilin fields) can be described by dissolution into the formation water alone, within a narrow pH range of 5–5.3 (Supplementary Figs 3–6).

Carbonate reservoir data from McElmo dome show a change in $CO_2/^3He$ ratio of over an order of magnitude, with invariant $\delta^{13}C(CO_2)$ (Fig. 3b). This pattern is repeated in the two other carbonate-dominated fields (Supplementary Figs 7, 8). Invariant $\delta^{13}C(CO_2)$ in these fields allows us to discount a two-process model of precipitation and dissolution such as at Bravo dome (Fig. 3a). These data are consistent with CO_2 dissolution only into formation water in the pH range of 5.4–5.8 (Fig. 3b and Supplementary Figs 7, 8), a value similar to the pH obtained for the siliciclastic reservoirs and to values observed (pH 5.7) in carbonate-mineral-buffered formation water observed in the recent CO_2 injection studies on CO_2 breakthrough[19] in the Frio formation, Texas.

On a reservoir-engineering timescale, the early stages of CO_2 injection can result in a drop in pH and dissolution of carbonate minerals into the formation water[18–20,22]. Any significant CO_2 contribution to the reservoir CO_2 phase from re-dissolution of carbonates would be ^3He free and would therefore perturb the correlation between $CO_2/^3He$ ratio and ^4He and ^{20}Ne. As there is a clear correlation between $CO_2/^3He$ ratio and ^4He in all fields and ^{20}Ne within the majority, we conclude that dissolution of carbonate minerals into the formation water cannot have had a major influence on $\delta^{13}C(CO_2)$ values. There is no evidence for any precipitation of CO_2 within the carbonate-dominated reservoirs, requiring that the dominant mechanism of reservoir CO_2 loss, accounting for up to 90%, is through dissolution into the formation water.

Even the most conservative model we have presented places an upper limit of approximately 18% on the CO_2 removed by precipitation, and then only in some samples, from all natural gas fields investigated in a variety of lithological settings. Precipitation of CO_2 over millennial timescales represents at most only a small subsurface trapping mechanism for CO_2, and only within siliciclastic lithologies. The dominant mechanism of CO_2 loss from most CO_2 natural gas fields can be accounted for through simple dissolution into the formation groundwater within a narrow pH window (pH 5–5.8). This study underscores the fact that understanding geological carbon storage requires careful investigation of existing geological and hydrogeological analogues that have naturally accumulated and stored CO_2 over timescales relevant to anthropogenic CO_2 storage facilities. We have also demonstrated a means of testing trapping and storage mechanisms through coupled measurements of noble gas and carbon isotopes in the context of the pH evolution of formation/reservoir water.

METHODS SUMMARY

Detailed descriptions of the sample collection and analysis procedures can be found in the original references[5,7,8,12,13,21]. In our calculations (Fig. 3 and Supplementary Figures) we use the highest $CO_2/^3He$ ratio measured in each field as a reference point to calculate the correlated reservoir $CO_2/^3He$ and $\delta^{13}C(CO_2)$ ratios as the CO_2 phase is removed by either precipitation or dissolution. We assume open system loss. In the case of precipitation there is zero ^3He loss from the CO_2 phase and $CO_2/^3He$ changes in proportion to the fraction of the remaining CO_2 phase. In the case of dissolution, the change in $CO_2/^3He$ ratio is calculated following the Rayleigh equation.

Changes in $\delta^{13}C(CO_2)$ are calculated using the Rayleigh fractionation equation expressed as $\delta^{13}C(CO_2) = \delta^{13}C(CO_2)_o + \varepsilon \ln f$ (ref. 23), where $\delta^{13}C(CO_2)_o$ is the original system value, f is the fraction of CO_2 remaining in the reservoir and ε is the carbon isotope fractionation, either for precipitation or for dissolution. Carbon isotope fractionation factors, α, are calculated as a function of temperature for

CO_2(g) precipitating to form $CaCO_3$(s), or dissolving to form either H_2CO_3(aq) or HCO_3^-(aq) (ref. 24). Because all the fractionations are small, the simplification can be made that $\varepsilon = 1{,}000\ln\alpha$ (ref. 25). For typical reservoir waters of pH 5–8, the contribution of CO_3^{2-}(aq) is negligible. Hence, for CO_2 dissolution, carbon isotope fractionation between the pool of dissolved inorganic carbon (DIC) and CO_2 gas used in the Rayleigh fractionation equation can be expressed as (ref. 23)

$$\varepsilon^{13}C_{DIC-CO_2(g)} = x(\varepsilon^{13}C_{H_2CO_3(aq)-CO_2(g)}) + (1-x)(\varepsilon^{13}C_{HCO_3^-(aq)-CO_2(g)})$$

where x is the proportion of CO_2(g) dissolving to H_2CO_3(aq) at the relevant pH[23].

Solubility as a function of temperature and salinity is given by the IUPAC solubility series for CO_2 (ref. 26) and in refs 27, 28 for He. The average well depth, reservoir pressure, temperature and salinity are presented in the Supplementary Information for each reservoir, with the corresponding Henry's law constants K_{He} and K_{CO_2} and fractionation factor ($1{,}000\ln\alpha$) for CO_2(g) forming H_2CO_3(aq), HCO_3^-(aq) and $CaCO_3$(s) (Supplementary Table 1).

Received 24 June 2008; accepted 22 January 2009.

1. Schrag, D. P. Preparing to capture carbon. *Science* **315,** 812–813 (2007).
2. Baines, S. J. & Worden, R. H. in *Geological Storage of Carbon Dioxide* (eds Baines, S. J. & Worden, R. H.) 1–6 (The Geological Society of London, 2004).
3. Gale, J. in *Geological Storage of Carbon Dioxide* (eds Baines, S. J. & Worden, R. H.) 7–15 (The Geological Society of London, 2004).
4. Bradshaw, J., Boreham, C. & La Pedalina, F. in *Proc. 7th Internat. Conf. Greenhouse Gas Control Technol. (GHGT-7)* (eds Rubin, E., Keith, D. & Gilboy, C.) 541–550 (Elsevier Science, 2004).
5. Ballentine, C. J., Schoell, M., Coleman, D. & Cain, B. A. 300-Myr-old magmatic CO_2 in natural gas reservoirs of the west Texas Permian basin. *Nature* **409,** 327–331 (2001).
6. Kintisch, E. The greening of synfuels. *Science* **320,** 306–308 (2008).
7. Gilfillan, S. M. V. *et al.* The noble gas geochemistry of natural CO_2 gas reservoirs from the Colorado Plateau and Rocky Mountain provinces, USA. *Geochim. Cosmochim. Acta* **72,** 1174–1198 (2008).
8. Sherwood Lollar, B., Ballentine, C. J. & O'Nions, R. K. The fate of mantle-derived carbon in a continental sedimentary basin: Integration of C/He relationships and stable isotope signatures. *Geochim. Cosmochim. Acta* **61,** 2295–2308 (1997).
9. Ballentine, C. J., Burgess, R. & Marty, B. in *Noble Gases in Geochemistry and Cosmochemistry* (eds Porcelli, D. R., Ballentine, C. J. & Weiler, R.) 539–614 (Geochemical Society and Mineralogical Society of America, 2002).
10. Cathles, L. M. & Schoell, M. Modeling CO_2 generation, migration and titration in sedimentary basins. *Geofluids* **7,** 441–450 (2007).
11. Baines, S. J. & Worden, R. H. in *Geological Storage of Carbon Dioxide* (eds Baines, S. J. & Worden, R. H.) 59–85 (The Geological Society of London, 2004).
12. Xu, S., Nakai, S., Wakita, H., Xu, Y. & Wang, X. Carbon isotopes of hydrocarbons and carbon dioxide in natural gases in China. *J. Asian Earth Sci.* **15,** 89–101 (1997).
13. Xu, S., Nakai, S., Wakita, H. & Wang, X. Mantle-derived noble gases in natural gases from Songliao Basin, China. *Geochim. Cosmochim. Acta* **59,** 4675–4683 (1995).
14. Ballentine, C. J. & Burnard, P. G. in *Noble Gases in Geochemistry and Cosmochemistry* (eds Porcelli, D. R., Ballentine, C. J. & Weiler, R.) 481–538 (Geochemical Society and Mineralogical Society of America, 2002).
15. Ballentine, C. J. & Sherwood Lollar, B. Regional groundwater focusing of nitrogen and noble gases into the Hugoton-Panhandle giant gas field, USA. *Geochim. Cosmochim. Acta* **66,** 2483–2497 (2002).
16. Torgersen, T. & Clarke, W. B. Helium accumulation in groundwater. I: An evaluation of sources and the continental flux of crustal ^{4}He in the Great Artesian Basin, Australia. *Geochim. Cosmochim. Acta* **49,** 1211–1218 (1985).
17. Broadhead, R. F. Natural accumulations of carbon dioxide in the New Mexico region - Where are they, how do they occur and what are the uses for CO_2? *Lite Geol.* **20,** 2–6 (1998).
18. Pearce, J. *et al.* Natural occurrences as analogues for the geochemical disposal of carbon dioxide. *Energy Convers. Manage.* **37,** 1123–1128 (1996).
19. Kharaka, Y. K. *et al.* Gas-water-rock interactions in Frio Formation following CO2 injection: Implications for the storage of greenhouse gases in sedimentary basins. *Geology* **34,** 577–580 (2006).
20. Knauss, K. G., Johnson, J. W. & Steefel, C. I. Evaluation of the impact of CO_2, co-contaminant gas, aqueous fluid and reservoir rock interactions on the geologic sequestration of CO_2. *Chem. Geol.* **217,** 339–350 (2005).
21. Xu, S., Shun'ichi, N., Wakita, H., Xu, Y. & Wang, X. Helium isotope compositions in sedimentary basins in China. *Appl. Geochem.* **10,** 643–656 (1995).
22. Worden, R. H. & Smith, L. K. in *Geological Storage of Carbon Dioxide* (eds Baines, S. J. & Worden, R. H.) 211–224 (The Geological Society of London, 2004).
23. Clark, I. D. & Fritz, P. *Environmental Isotopes in Hydrology* 55–61 (CRC, 1997).
24. Deines, P., Langmuir, D. & Harmon, R. S. Stable carbon isotopes and the existence of a gas phase in the evolution of carbonate groundwaters. *Geochim. Cosmochim. Acta* **38,** 1147–1184 (1974).
25. Fritz, P. & Fontes, J. C. *Handbook of Environmental Isotope Geochemistry* Vol. 1, 1–19 (Elsevier, 1980).
26. Scharlin, P. & Cargill, R. W. *Carbon Dioxide in Water and Aqueous Electrolyte Solutions* (Solubility Data Series Vol. 62, IUPAC, 1996).
27. Crovetto, R., Fernandez-Prini, R. & Laura Japas, M. Solubilities of inert gases and methane in H_2O and in D_2O in the temperature range of 300 to 600K. *J. Chem. Phys.* **76,** 1077–1086 (1982).
28. Smith, S. P. Noble gas solubility in water at high temperature. *Eos* **66,** 397 (1985).
29. Sherwood Lollar, B., O'Nions, R. K. & Ballentine, C. J. Helium and neon isotope systematics in carbon dioxide-rich and hydrocarbon-rich gas reservoirs. *Geochim. Cosmochim. Acta* **58,** 5279–5290 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements S.M.V.G. was supported by a Natural Environmental Research Council (NERC)-funded PhD studentship in Manchester and a NERC-funded postdoctoral position, grant NE/C516479/1 in Edinburgh and Glasgow, and UK Energy Research Centre grant NE/C513169/1. Manchester work was further partly funded by NERC grants NE/D004292 and NE/F002823. Toronto work was further partly funded by an Natural Sciences and Engineering Research Council of Canada Discovery grant to B.S.L. We thank the field operators for permission to sample the US gas reservoirs and support in the field, particularly L. Nugent (Sheep Mountain), T. Muhic and D. Miller and G. Grove (McCallum dome) and T. White (St Johns dome). S.M.V.G. would like to thank R. S. Haszeldine and Z. Shipton for supporting this work. Review by R. H. Worden is appreciated.

Author Contributions S.M.V.G., C.J.B. and B.S.L. designed the study, analysed the samples, interpreted the data and wrote the paper. G.H., D.B., Z.D., Z.Z. and G.L.-C. assisted with sample analysis and interpretation of the data. S.S., M.S. and M.C. assisted with sample collection and provided comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.M.V.G. (stuart.gilfillan@ed.ac.uk).

LETTERS

Petrological evidence for secular cooling in mantle plumes

Claude Herzberg[1] & Esteban Gazel[1]

Geological mapping and geochronological studies have shown much lower eruption rates for ocean island basalts (OIBs) in comparison with those of lavas from large igneous provinces (LIPs) such as oceanic plateaux and continental flood provinces[1]. However, a quantitative petrological comparison has never been made between mantle source temperature and the extent of melting for OIB and LIP sources. Here we show that the MgO and FeO contents of Galapagos-related lavas and their primary magmas have decreased since the Cretaceous period. From petrological modelling[2], we infer that these changes reflect a cooling of the Galapagos mantle plume from a potential temperature of 1,560–1,620 °C in the Cretaceous to 1,500 °C at present. Iceland also exhibits secular cooling, in agreement with previous studies[3,4]. Our work provides quantitative petrological evidence that, in general, mantle plumes for LIPs with Palaeocene–Permian ages were hotter and melted more extensively than plumes of more modern ocean islands. We interpret this to reflect episodic flow from lower-mantle domains that are lithologically and geochemically heterogeneous.

Extensive outcrops of basalt, picrite, and sometimes komatiite ~65–95 Myr old occupy portions of the Caribbean LIP (CLIP). It has been suggested[5] that they were produced by melting in the Galapagos mantle plume, and this is consistent with isotopic and geochemical similarities with lavas from the present-day Galapagos hotspot[6]. A Galapagos link for rocks in South American oceanic complexes is more controversial. Basalts, picrites, and komatiites from Gorgona Island, Columbia, were originally considered part of the CLIP[7,8]. However, other studies[9] suggest Gorgona and other South American complexes were once part of a separate oceanic plateau related to Salas y Gomez Island, Chile, or some other hotspot (Supplementary Information).

The lowest FeO contents are mostly found in lavas 0–13 Myr old from the present-day Galapagos archipelago and the Carnegie ridge and Cocos ridge hotspot tracks (Fig. 1a). FeO contents are highest for Gorgona komatiites and intermediate for all other lavas. When olivine is the sole crystallizing phase, lavas with higher FeO contents can be differentiated from peridotite-source primary magmas with higher FeO and MgO contents[2,3,10–12] (Fig. 1a). A primary magma is a partial melt of the mantle formed, in most cases, by the mixing of small melt droplets that are separated from the remainder of the solid residue[2,3,10,11]. Addition or subtraction of olivine from a primary magma will produce lavas having higher or lower MgO contents, respectively, with minor change in FeO content. We simulated this and reconstructed the primary magma compositions using the PRIMELT2 model of ref. 2 (Methods Summary). Our results are given in Supplementary Information and Fig. 1a.

The MgO content of a volatile-deficient primary magma is positively correlated with the temperature of the mantle[2,3,10–12]. It provides a petrological record of mantle potential temperature, T_P, which is the temperature that the solid adiabatically convecting mantle would attain if it could reach the surface without melting[13]. Using the relationship $T_P = 1{,}463 + 12.74\mathrm{MgO} - 2{,}924/\mathrm{MgO}$ (refs 2, 3; here MgO is measured in weight per cent and T_P is given in degrees Celsius), we can now readily calculate how hot the mantle had to be to yield the primary magma compositions given in Fig. 1a. Our results are shown in Figs 1b and 2. For the present-day Galapagos plume, T_P ranges from 1,400 to 1,500 °C (ref. 2), similar to the T_P range of 1,440–1,500 °C recorded for lavas from the Cocos and Carnegie ridges. Older lavas were hotter. Those from the CLIP and accreted tracks with ages of 65–95 Myr have a T_P range of 1,500 to 1,560 °C, and up to 1,620 °C if Gorgona lavas were part of the CLIP. This is petrological evidence for secular cooling of the Galapagos plume.

The MgO content of an accumulated fractional melt does not change substantially as melt fraction increases during decompression[2,3,11]. The adiabatic temperature–pressure melting path is approximately coincident with the olivine liquidus, which can be calculated using $T_{OL} = 935 + 33\mathrm{MgO} - 0.37(\mathrm{MgO})^2 + 54P - 2P^2$, where T_{OL} and MgO are measured as above and pressure, P, is measured in gigapascals[2,11]. Using final melting pressures and the MgO contents of primary magmas in this equation (Fig. 1a), a synthetic adiabatic melting path can be obtained (Fig. 1b). The majority of lavas from the present-day Galapagos plume formed in a column where melting ended at >2 GPa, and this pressure is highly variable. Melting ended at much lower pressures for lavas from the Cocos and Carnegie ridges, consistent with the channelling of the Galapagos plume to locations of thinner lithosphere. Low pressures of final melting are also inferred for many older CLIP lavas, indicating the possible involvement of thin lithosphere associated with ocean ridges.

We now provide petrological evidence for secular cooling in other areas. Results given in Supplementary Information and Fig. 2 illustrate that LIPs dating from the Palaeocene epoch and earlier were formed by mantle sources that were generally hotter than present-day ocean islands. However, there are several important exceptions. First, Hawaii is the ocean island that is most similar to a LIP, in that it has a maximum T_P of 1,600 °C. It is only surpassed by rocks from the North Atlantic igneous province, the Deccan Traps and the CLIP if we include Gorgona. Second, T_P for the Central Atlantic magmatic province (CAMP) is notably different from all other LIPs in being cool (Fig. 2). The T_P excess of ~100 °C for the CAMP is consistent with model temperatures[14] that can arise from an internally heated mantle capped by Pangaea[14,15]. This is evidence indicating that continental insulation is not capable of the producing LIPs with the much higher values of T_P (Fig. 2).

Noteworthy is the wide range of primary magma compositions and inferred mantle potential temperatures for each LIP and ocean island occurrence (Fig. 2). These ranges have been interpreted as originating from a hotspot, a spatially localized source of heat and magmatism restricted in time[2]. Primary magmas are tapped from both the hot axis and the cool periphery of the plume as illustrated

[1]Department of Earth and Planetary Sciences, Rutgers University, 610 Taylor Road, Piscataway, New Jersey 08854-8066, USA.

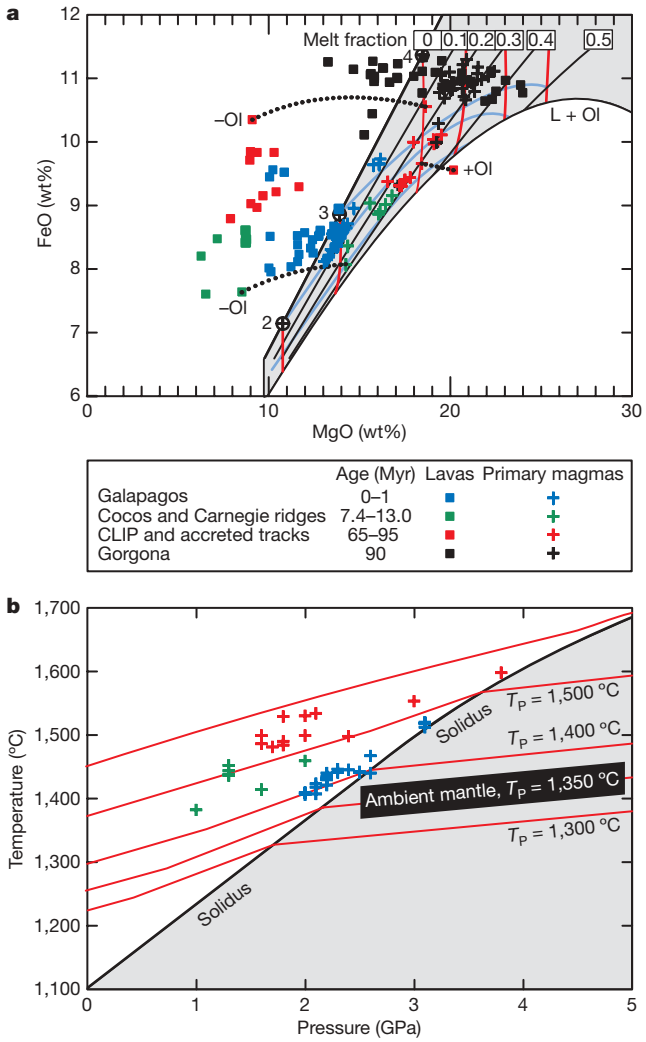


Figure 1 | Compositions and inferred temperature–pressure conditions of melting for Galapagos-related magmatism. **a**, FeO and MgO contents of lavas and calculated primary magmas from the present-day Galapagos hotspot, the Cocos and Carnegie ridges, old accreted Galapagos tracks and the CLIP. Lavas from Gorgona are plotted separately because is not clear whether they were part of the CLIP[7,8] or some other oceanic complex[9]. Lines of filled circles identify liquid compositions that result from olivine addition to (+) and subtraction from (−) specific lava compositions. Primary magma compositions were computed using PRIMELT2[2]. Primary magmas of fertile peridotite KR-4003 are plotted within the grey-coloured area[2]. The intersection of a red line (initial melting pressure) and a blue line (final melting pressure) identifies the composition of an accumulated fractional melt at the pressure of initial and final melting. Individual lavas and their sources from which primary magmas are calculated are identified in Supplementary Table 1. Pressure is indicated in gigapascals by the circled crosses, as shown. L, liquid; Ol, olivine. **b**, Inferred temperatures and pressures at which fractional melting terminated (Methods). Red lines, adiabatic melting paths[11]. Gorgona komatiites probably formed from a more depleted peridotite source[3], and solutions are not provided.

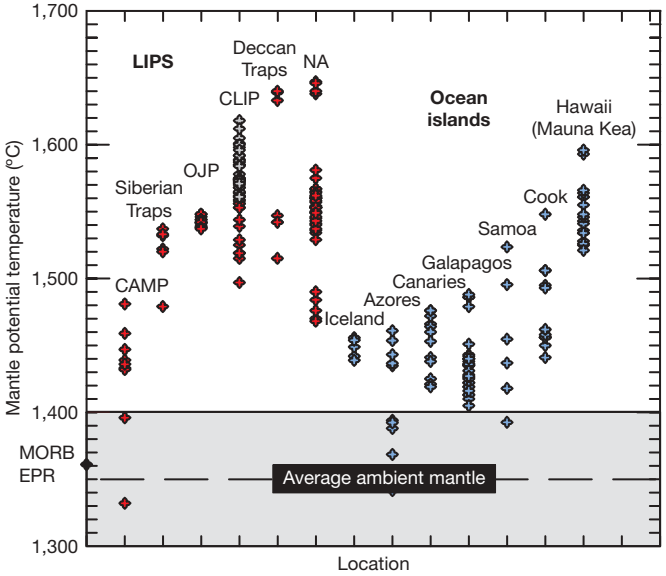


Figure 2 | Mantle potential temperatures inferred for lavas from some LIPs and ocean islands. T_P has been computed from primary magma MgO content using PRIMELT2[2]. Data sources and calculated T_P values for ocean islands and LIPS are given in Supplementary Information. CLIP results are for rocks with ages ≥65 Myr, and include old accreted Galapagos tracks. Gorgona data is shown separately using grey crosses. Galapagos results are from lavas within the archipelago. OJP, primary magmas for lavas from the Ontong Java plateau; NA, primary magmas for Palaeocene lavas from the North Atlantic igneous province found in East and West Greenland; CAMP, primary magmas for the Central Atlantic magmatic province; MORB, mid-ocean-ridge basalt; EPR, East Pacific Rise.

in Fig. 3. The T_P maximum of 1,500 °C for Galapagos is characteristic of the plume axis. The lower end of the Galapagos range approaches 1,350 ± 50 °C, a T_P value for ambient mantle[3,10,16,17] necessary for the production of MORB with 10–13 wt% MgO. What is particularly relevant for our purposes is that there is a decrease in T_P maxima from 1,560–1620 °C for rocks 65–95 Myr old to 1,500 °C at present (Fig. 2). The exact form of the secular cooling curve depends on whether the Gorgona komatiites were produced by the Galapagos plume or another (Supplementary Information).

Melt fractions computed from PRIMELT2 are generally higher for LIPs than for ocean islands (Fig. 4), consistent with suggestions of higher eruption rates[1]. The high melt fractions, high mantle potential temperatures and vast areas of magmatism associated with the largest LIPs are all consistent with formation in mantle plume heads[1] (but note the possible CAMP exception). By contrast with LIPs, many ocean islands display melt fractions that must be lower than ~0.05 (Fig. 4b). These are often readily characterized by very low SiO_2, high CaO and high lithophile trace-element abundances in OIBs owing to low-degree melting of carbonated peridotite[2,18]. Low-melt-fraction, CO_2-rich OIBs are abundant in the Azores, the Canary Islands, Cape Verde, the Cook–Austral chain, the Marquesas Islands, the Pitcairn–Gambier chain, St Helena, Samoa and the Society Islands, and many other ocean islands (see, for example, the Geochemistry of Rocks of the Oceans and Continents database (http://georoc.mpch-mainz.gwdg.de/georoc/) and Supplementary Information). Even more of this OIB-type melt is likely to metasomatize the mantle rather than erupt. The melt-fraction frequency spectrum for OIB in Fig. 4b is

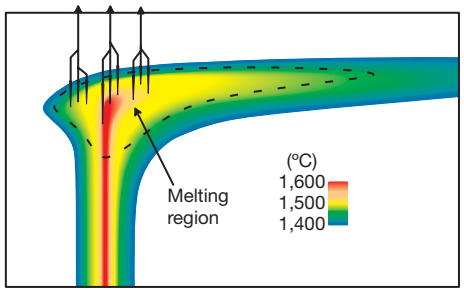


Figure 3 | A generic model for interpreting the spatial localization of petrological variability. The model provides an interpretation of primary magmas with highly variable compositions, inferred mantle potential temperatures and melt fractions. This is the mantle plume model in which hot primary magmas originate from the axis and cooler primary magmas originate from the periphery. The colour bar indicates mantle potential temperature.

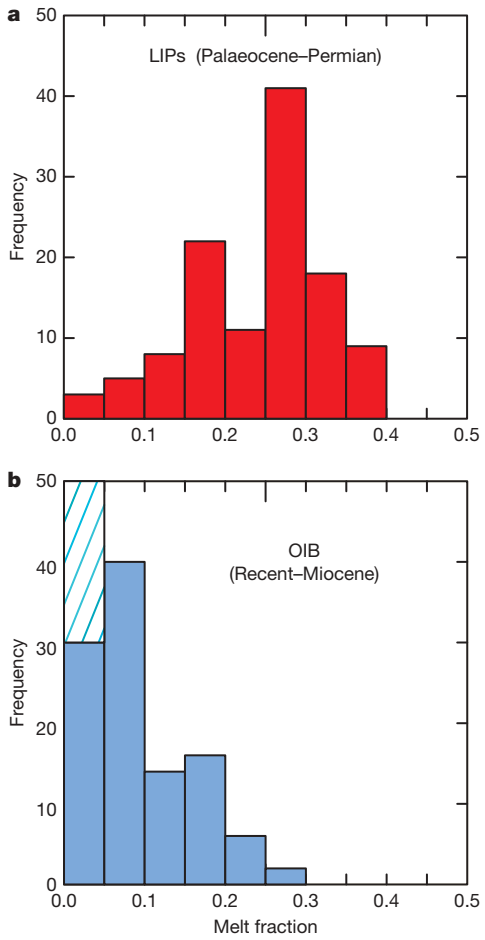


Figure 4 | Melt fractions inferred for lavas for some LIPs and ocean islands. Melt fractions have been computed using PRIMELT2[2], and refer to the total melt fraction with respect to source mass for accumulated fractional melting of fertile peridotite. **a**, LIPs. Data sources for model primary magmas for LIPS are given in Supplementary Information. **b**, OIB. Solid blue bars indicate primary magma solutions from ocean islands (see ref. 2 and Supplementary Information). The hatched region indicates an abundance of OIB melted from volatile-enriched sources at very low melt fractions; these are generally more abundant than volatile-deficient lavas, and cannot be modelled with PRIMELT2 (see fig. 11 in ref. 2). Frequency is the number of primary magma solutions.

therefore likely to be exponential in form. Results for these OIB occurrences are interpreted as the transport of low-melt-fraction magmas from the cool plume peripheries and high-melt-fraction magmas from the hotter plume axes (Fig. 3). However, it has been proposed that low-melt-fraction OIB can also form without a plume by volatile-induced melting of ambient mantle and transport through lithospheric fractures[19]. This suggestion is fully consistent with experiments[18] and PRIMELT2[2] modelling. Both plume and non-plume origins are indicated for ocean islands.

A very high cooling rate is inferred for the Icelandic plume. Most Palaeocene lavas with ~60-Myr pre-breakup ages[20] from East and West Greenland have T_P maxima of ~1,550–1,570 °C, similar to the CLIP, and crystallized from primary magmas with 18–20 wt% MgO. Our model primary magmas are in excellent agreement with many previous estimates[3,4,11,21], although we obtained T_P values as high as 1,650 °C (Fig. 2). A spread of ~200 °C in T_P and melt fractions in the range 0.05−0.37 have been recorded in East Greenland lavas (Supplementary Information) from a restricted area close to the Tertiary Icelandic hotspot track[22]. These ranges are an expected consequence of the tapping of primary magmas from a mantle plume (Fig. 3). A T_P value as low as 1,460 °C has been obtained from lavas with ~55-Myr syn-breakup ages from the seaward-dipping reflector sequence, similar to present-day Iceland[2,3,23] (Fig. 2). Our work indicates that T_P decreased from the range 1,550–1,650 °C to 1,460 °C in about 5 Myr, in agreement with estimates in ref. 4. The T_P value for the Icelandic plume appears unchanged at about 1,460 °C from 55 Myr ago to the present, and is now in a comparatively steady state. The early rapid secular cooling of the Icelandic plume is much greater than that seen for the Galapagos, although more work is needed to fill the gap in the Galapagos data (Supplementary Information). We also acknowledge that an Icelandic plume cooling curve is compromised by an absence of data from the Greenland–Iceland and Iceland–Faeroes ridges with ~15–50-Myr ages.

Our work provides petrological evidence that mantle plumes for LIPs with Palaeocene–Permian ages were hotter and melted more extensively than plumes of more modern ocean islands. One interpretation is that LIPs melted from large plume heads and OIBs melted from thin plume conduits[1], and cooling is more effective in the latter. Indeed, there is now an important literature on lithosphere and asthenosphere cooling of mantle plumes[24,25]. However, this explanation fails to explain why hot LIPs such as those in Fig. 2 are not erupting today.

Numerical and laboratory simulations show that mantle flow can be episodic where there are thermal and compositional components to buoyancy[25–27]. Mantle plumes with these characteristics might originate in lower-mantle domains where shear-wave velocities are low and bulk density is intrinsically high[28,29]. Subduction can contribute to high silica content[30], and iron content that is both high[30] and low in these domains (Methods), and mixing may yield heterogeneities on a range of length scales. Plumes may randomly sample this complexity, or lighter components may preferentially separate from more dense lithologies that stay behind. Although progress is being made on identifying peridotite and subducted crustal source lithologies from the compositions of lavas, inferring iron content is a much more difficult problem (Methods). Nevertheless, we are optimistic that integrated petrological and deep-mantle studies can provide a better picture of the birth–life–death cycle of mantle plumes.

METHODS SUMMARY

Primary magma compositions, mantle potential temperatures and source melt fractions were calculated from primitive whole-rock compositions using PRIMELT2 spreadsheet software[2]. A detailed discussion of the method is given elsewhere[2,3,11]. The algorithm calculates the primary magma composition for a primitive lava by determining the variable amounts of olivine that were added or subtracted.

PRIMELT2 was calibrated on the basis of experiments on fertile peridotite with 8 wt% FeO, and all calculated primary magma compositions were assumed to have been derived by fractional melting. For each primary magma, it provided the olivine liquidus temperature, T_{OL}, at 1 atm and the mantle potential temperature, T_P. As both T_{OL} and T_P depend on the MgO content of the primary magma[3], the accuracy of the former is a guide to the precision of the latter. For any specific peridotite composition, the uncertainty in T_{OL} is ±31 °C at the 2σ confidence level[3]. Uncertainties in the FeO content of peridotite can propagate to an uncertainty of ±50–70 °C in T_P (Methods). Uncertainties in all other major elements for fertile peridotite do not propagate to significant variations in melt fraction and mantle potential temperature[2,11]. Melting of depleted peridotite propagates to calculated melt fractions that are too high, but with a negligible error in mantle potential temperature[2,3,11].

We used PRIMELT2 to identify magmas generated from pyroxenite sources, and excluded them. Magmas that have been degassed from CO_2-rich sources were identified and similarly excluded. Fe_2O_3 content was calculated using $Fe_2O_3/TiO_2 = 0.5$, a reduced mode, on the basis of MORB-like FeO enrichment for most LIPs[2]. Lavas that had experienced plagioclase and/or clinopyroxene fractionation were excluded from this analysis.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 September 2008; accepted 28 January 2009.

1. Richards, M. A., Duncan, R. A. & Courtillot, V. E. Flood basalts and hot-spot tracks: plume heads and tails. *Science* **246**, 103–107 (1989).
2. Herzberg, C. & Asimow, P. D. Petrology of some oceanic island basalts: PRIMELT2.XLS software for primary magma calculation. *Geochem. Geophys. Geosyst.* **9**, doi:10.1029/2008GC002057 (2008).

3. Herzberg, C. *et al.* Temperatures in ambient mantle and plumes: constraints from basalts, picrites and komatiites. *Geochem. Geophys. Geosyst.* **8,** doi:10.1029GC001390 (2007).
4. Armitage, J. J., Henstock, T. J., Minshull, T. A. & Hopper, J. R. Modelling the composition of melts formed during continental breakup of the Southeast Greenland margin. *Earth Planet. Sci. Lett.* **269,** 248–258 (2008).
5. Duncan, R. A. & Hargraves, R. B. Plate tectonic evolution of the Caribbean region in the mantle reference frame. *Bull. Geol. Soc. Am.* **162,** 81–93 (1984).
6. Hoernle, K., Hauff, F. & van den Bogaard, P. 70 m.y. history (139–69 Ma) for the Caribbean large igneous province. *Geology* **32,** 697–700 (2004).
7. Storey, M., Mahoney, J. J., Kroenke, L. W. & Saunders, A. D. Are oceanic plateaus sites for komatiite formation? *Geology* **19,** 376–379 (1991).
8. Kerr, A. C. *et al.* The petrogenesis of Gorgona komatiites, picrites and basalts: new field, petrographic and geochemical constrains. *Lithos* **37,** 245–260 (1996).
9. Kerr, A. C. & Tarney, J. Tectonic evolution of the Caribbean and northwestern South America: The case for accretion of two Late Cretaceous oceanic plateaus. *Geology* **33,** 269–272 (2005).
10. Langmuir, C. H., Klein, E. M. & Plank, T. in *Mantle Flow and Melt Generation at Mid-Ocean Ridges* (eds Morgan, J. P., Blackman, D. K. & Sinton J. M.) 183–280 (Geophys. Monogr. Ser. 71, American Geophysical Union, 1992).
11. Herzberg, C. & O'Hara, M. J. Plume-associated ultramafic magmas of Phanerozoic age. *J. Petrol.* **43,** 1857–1883 (2002).
12. Putirka, K. D. Mantle potential temperatures at Hawaii, Iceland, and the mid-ocean ridge system, as inferred from olivine phenocrysts: evidence for thermally driven mantle plumes. *Geochem. Geophys. Geosyst.* **6,** doi:10.1029/2005GC000915 (2005).
13. McKenzie, D. & Bickle, M. J. The volume and composition of melt generated by extension of the lithosphere. *J. Petrol.* **29,** 625–679 (1988).
14. Coltice, N., Phillips, B. R., Bertrand, H., Richard, Y. & Rey, P. Global warming of the mantle at the origin of flood basalts over supercontinents. *Geology* **35,** 391–394 (2007).
15. Anderson, D. L. Hotspots, polar wander, Mesozoic convection and the geoid. *Nature* **297,** 391–393 (1982).
16. McKenzie, D., Jackson, J. & Priestley, K. Thermal structure of oceanic and continental lithosphere. *Earth Planet. Sci. Lett.* **233,** 337–349 (2005).
17. Courtier, A. M. *et al.* Correlation of seismic and petrological thermometers suggests deep thermal anomalies beneath hotspots. *Earth Planet. Sci. Lett.* **264,** 308–316 (2007).
18. Dasgupta, R., Hirschmann, M. M. & Smith, N. D. Partial melting experiments on peridotite + CO_2 at 3 GPa and genesis of alkalic ocean island basalts. *J. Petrol.* **48,** 2093–2124 (2007).
19. Hirano, N. *et al.* Volcanism in response to plate flexure. *Science* **313,** 1426–1428 (2006).
20. Storey, M., Ducan, R. A. & Tegner, C. Timing and duration of volcanism in the North Atlantic igneous province: implications for geodynamics and links to the Iceland hotspot. *Chem. Geol.* **241,** 264–281 (2007).
21. Holm, P. M. *et al.* The tertiary picrites of West Greenland: contributions from 'Icelandic' and other sources. *Earth Planet. Sci. Lett.* **115,** 227–244 (1993).
22. Saunders, A. D., Fitton, J. G., Kerr, A. C., Norry, M. J. & Kent, R. W. in *Large Igneous Provinces: Continental, Oceanic, and Planetary Flood Volcanism* (eds Mahoney, J. J. & Coffin, M. J.) 45–93 (Geophys. Monogr. Ser. 100, American Geophysical Union, 1997).
23. Slater, L., McKenzie, D., Grönvold, K. & Shimizu, N. Melt generation and movement beneath Theistareykir, NE Iceland. *J. Petrol.* **42,** 321–354 (2001).
24. Sleep, N. Channeling at the base of the lithosphere during the lateral flow of plume material beneath flow line hot spots. *Geochem. Geophys. Geosyst.* **9,** doi:10.1029/2008GC002090 (2008).
25. Kumagai, I., Davaille, A., Kurita, K. & Stutzmann, E. Mantle plumes: thin, fat, successful, or failing? Constraints to explain hot spot volcanism through time and space. *Geophys. Res. Lett.* **35,** doi:10.1029/2005GL035079 (2008).
26. Farnetani, C. G. & Samuel, H. Beyond the thermal plume paradigm. *Geophys. Res. Lett.* **32,** doi:10.1029/2005GL022360 (2005).
27. Lin, S.-C. & van Keken, P. E. Multiple volcanic episodes of flood basalts caused by thermochemical mantle plumes. *Nature* **436,** 250–252 (2005).
28. Garnero, E. J. & McNamara, A. K. Structure and dynamics of Earth's lower mantle. *Science* **320,** 626–628 (2008).
29. Burke, K., Steinberger, B., Torsvik, T. H. & Smethurst, M. A. Plume generation zones at the margins of large low shear velocity provinces on the core–mantle boundary. *Earth Planet. Sci. Lett.* **265,** 49–60 (2008).
30. Trampert, J., Deschamps, F., Resovsky, J. & Yuen, D. Probabilistic tomography maps chemical heterogeneities throughout the lower mantle. *Science* **306,** 853–856 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to N. Sleep and A. Kerr for reviews, and to C. Class, M. Hirschmann, P. Asimow, M. Humayun and K. Hoernle for discussions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.H. (herzberg@rci.rutgers.edu).

METHODS

We assumed that all OIB and LIP lavas melted from peridotite with 8.0 wt% FeO, which is the average for natural fertile and depleted peridotite occurrences[11]. We acknowledge, however, that mantle plume sources might differ if they originated in chemically unusual lower-mantle domains where shear-wave velocities are low and density is intrinsically high[28]. These domains may contain subducted oceanic crustal rocks of Archaean and Proterozoic ages, which are iron-rich picrites[31] and which might have reacted with host peridotite to produce a variety of iron-rich peridotite and pyroxenite lithologies as inferred from seismic and geodynamic data[30,32]. Iron-rich crust would have left behind complementary iron-poor peridotite residues with FeO contents <8.0% (ref. 31); that which did not construct cratonic lithospheric mantle[31] could have been subducted to yield lower-mantle domains that are low in FeO. There is likely to be substantial heterogeneity in iron on a scale that is too fine to be resolved using seismic data.

Progress has been made on identifying peridotite and pyroxenite source lithologies from the compositions of lavas[33,34], and this is encoded in PRIMELT2[2]. However, inferring the iron content of a source from a lava composition is a much more difficult problem. If some OIB and LIPs melted from iron-rich peridotite with 9 wt% FeO, for example, model primary magmas will be too high in MgO[2,11], and the mantle potential temperatures summarized in Fig. 2 will be 50–70 °C too high. This is strictly an artefact of the computational method for primary magma calculation, and is not to be confused with higher mantle potential temperatures that are needed to make iron-rich mantle buoyant. Greater iron enrichment is not likely, as it would propagate to lava SiO_2 contents that are lower than observed, on the basis of experimental results of Kushiro[35]. For iron-poor peridotite with 7 wt% FeO, potential temperatures will be too low by about 70 °C. There is little else we can do at present other than acknowledge the potential importance of iron variability in lower-mantle plume sources[30,32].

We have assumed that primary magmas are formed by accumulated fractional melting[2,3,10,11]. The initial melting pressure, P_i, and final melting pressure, P_f, are indicated in Fig. 1a by the red and blue lines, respectively. These have been calculated by forward simulations of fractional melting of fertile peridotite[11,31]. The final melting pressure is useful because it permits the construction of a synthetic temperature–pressure adiabatic melting path (Fig. 1b). The final melting pressure can be inferred by simply plotting FeO content and MgO content for a PRIMELT2 primary magma in Fig. 1a and interpolating using the blue lines. Alternatively, the final melting pressure can be calculated using the following equations. For primary magmas with <15 wt% MgO

$$P_{1f} = a + b\mathrm{FeO} + c(\mathrm{FeO})^2$$

Here FeO is the weight per cent of iron in the primary magma and *a*, *b*, and *c* are variables that depend on the MgO content of the primary magma:

$$a = -196.4 + 2.942\mathrm{MgO} + 430/\mathrm{MgO}$$

$$b = 17.7 - 0.444\mathrm{MgO} + 228/\mathrm{MgO}$$

$$c = 2.2 - 0.047\mathrm{MgO} - 42.78/\mathrm{MgO}$$

For primary magmas with 15% > MgO < 20%, the appropriate pressure to use is

$$P_{2f} = P_{1f} - 10.96 + 0.67\mathrm{MgO}$$

The difference between calculated P_f values and those indicated in Fig. 1a by the blue lines is ± 0.28 GPa (2σ). Complex changes in phase equilibria will probably restrict pressure inferences for other OIB and LIP primary magmas to MgO < 20% and $P_f < 3.5$ GPa, similar to those for Galapagos and CLIP primary magmas.

Initial melting for garnet peridotite in the 2.7 GPa< P_i < 7 GPa range can be inferred by simply plotting FeO content and MgO content for a PRIMELT2 primary magma in Fig. 1a and interpolation using the red lines. Alternatively, they can be calculated from PRIMELT2 solutions for primary magma MgO contents using the equation

$$P_i = 11.248\mathrm{MgO} - 13{,}700(1/\mathrm{MgO})^3 - 8.13(\ln \mathrm{MgO})^3$$

where the difference between calculated P_i values and those indicated in Fig. 1a by the red lines is ± 0.20 GPa (2σ).

31. Herzberg, C. Geodynamic information in peridotite petrology. *J. Petrol.* **45,** 2507–2530 (2004).
32. Forte, A. M. & Mitrovica, J. X. Deep-mantle high-viscosity flow and thermochemical structure inferred from seismic and geodynamic data. *Nature* **410,** 1049–1056 (2001).
33. Sobolev, A. V., Hofmann, A. W., Sobolev, S. V. & Nikogosian, I. K. An olivine-free mantle source of Hawaiian shield basalts. *Nature* **434,** 590–597 (2005).
34. Herzberg, C. Petrology and thermal structure of the Hawaiian plume from Mauna Kea volcano. *Nature* **444,** 605–609 (2006).
35. Kushiro, I. in *Earth Processes: Reading the Isotopic Code* (eds Basu, A. & Hart, S.) 109–122 (Geophys. Monogr. Ser. 95, American Geophysical Union, 1996).

LETTERS

Initial community evenness favours functionality under selective stress

Lieven Wittebolle[1]*, Massimo Marzorati[1]*, Lieven Clement[2], Annalisa Balloi[4], Daniele Daffonchio[4], Kim Heylen[3], Paul De Vos[3], Willy Verstraete[1] & Nico Boon[1]

Owing to the present global biodiversity crisis, the biodiversity–stability relationship and the effect of biodiversity on ecosystem functioning have become major topics in ecology[1–3]. Biodiversity is a complex term that includes taxonomic, functional, spatial and temporal aspects of organismic diversity, with species richness (the number of species) and evenness (the relative abundance of species) considered among the most important measures[4,5]. With few exceptions (see, for example, ref. 6), the majority of studies of biodiversity-functioning and biodiversity–stability theory have predominantly examined richness[7–11]. Here we show, using microbial microcosms, that initial community evenness is a key factor in preserving the functional stability of an ecosystem. Using experimental manipulations of both richness and initial evenness in microcosms with denitrifying bacterial communities, we found that the stability of the net ecosystem denitrification in the face of salinity stress was strongly influenced by the initial evenness of the community. Therefore, when communities are highly uneven, or there is extreme dominance by one or a few species, their functioning is less resistant to environmental stress. Further unravelling how evenness influences ecosystem processes in natural and humanized environments constitutes a major future conceptual challenge.

Several components of biodiversity, such as species and functional group richness, have been shown to influence ecosystem function and stability significantly[3,12]. Species evenness has similarly been shown to influence community dynamics[13] and be an important element in managing invasions and production in managed ecosystems[14–16]. However, the influence of species evenness on the stability of ecosystem functioning remains unknown. Theoretically, evenness could strongly influence the stability of ecosystem functioning. For example, in a community where species are functionally redundant (that is, most contribute to the ecosystem function of interest), if initial evenness is high then the probability that a species tolerant to a perturbation is present is higher than when evenness is low. When evenness is low, meaning that the community is dominated by one or a few species, resistance to the perturbation will only occur if the dominant species are tolerant to the perturbation.

To test the relationship between initial community evenness and functionality, we used microcosm tests with denitrifying bacterial model communities. These are tools well suited to addressing ecological questions, as they can be maintained under simplified and defined conditions[17–19]. In addition, denitrifier models are good for investigating the value of microbial biodiversity in ecosystem functioning, owing to the wide range of physiological properties in this functional group of bacteria[20]. Different levels of initial evenness were assembled by, in each mixture, using eighteen different denitrifying species from four different phyla (Supplementary Table 3). We acknowledge that the degree of evenness will probably change during the course of the experiment. However, our hypothesis aims to test the response of the initial community evenness and how this translates itself into functional stability, regardless of any further shifts in community structure. A total of 1,260 microcosms, all with the same richness, were set up, incubated for 20 h under three distinct conditions (no stress, low temperature and salt stress), and related to the stability of the net ecosystem denitrification as a measure of the ecosystem functionality. All selected denitrifying species had similar activity response ranges, in order that all could contribute to ecosystem productivity. They represented an average range of richness broad enough to ensure good functionality[7]. Varying evenness without changing richness decreases the confounding of diversity by species identity[2,5].

Lorenz curves were used to assess initial community evenness visually. The Gini coefficient (ranging from zero to one) is a single value that describes a specific degree of evenness (Supplementary Fig. 3), measuring the normalized area between a given Lorenz curve and the perfect evenness line. The higher the Gini coefficient, the more uneven a community is. Lorenz curves of all 1,260 microcosms showed that almost the entire evenness range was sampled (Fig. 1). The net ecosystem denitrification of the investigated microbial

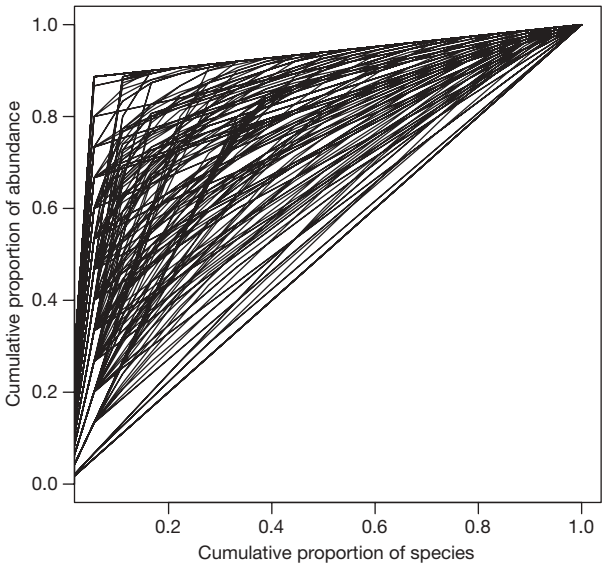


Figure 1 | Lorenz curves used in the experiment. The curves span the entire region between perfect evenness and high dominance.

[1]LabMET, Laboratory of Microbial Ecology & Technology, [2]BIOSTAT, Department of Applied Mathematics, Biometrics and Process Control, [3]LM-UGent, Laboratory of Microbiology, Department of Biochemistry, Physiology and Microbiology, Ghent University, B-9000 Ghent, Belgium. [4]DISTAM, Dipartimento di Scienze e Tecnologie Alimentari e Microbiologiche, Università degli Studi di Milano, 20133 Milan, Italy.
*These authors contributed equally to this work.

Table 1 | Linear models estimating the effects of various factors on the denitrification functionality

Step	Model	Residual d.f.	Residual SS	Treatment d.f.	Treatment SS	AIC
0	Intercept	1,439	29.06	—	—	−1,529.90
1	$0 + P + R + C + B + I$	1,398	21.43	41	7.63	−1,886.15
2	$1 + S$	1,396	13.44	2	7.99	−2,554.30
3	$2 + S{:}P$	1,388	10.25	8	3.19	−2,928.20
4	$3 + S{:}I$	1,352	7.96	36	2.29	−3,325.30
5	$4 + G^2 + S{:}G^2$	1,349	7.14	3	0.82	−3,370.40
6	$5 + B{:}S$	1,347	6.65	2	0.49	−3,469.80
7	$6 + S{:}R$	1,333	6.46	14	0.19	−3,484.20
8	$7 + S{:}C$	1,311	6.24	22	0.22	−3,489.40
9	$8 + X + S{:}X$	1,308	6.23	3	0.01	−3,485.60

The linear models describe the effects of stress, *S*, identity of the dominant species, *I*, Gini coefficient, *G*, and the relative abundance of the dominant species, *X*, on the functionality. The models also allow corrections for experiment effects, *P*, row effects, *R*, column effects, *C*, and the negative controls, *B*. Interactions are indicated using colons. At each step (1–9), terms were added to the model. The residual degrees of freedom (d.f.) and sum of squares (SS) are given. The treatment degrees of freedom and sum of squares only apply to the term that was added to the model. The Akaike information criterion[30] (AIC) was calculated for each model; a lower AIC indicates an improved model.

communities was expressed by the difference between the nitrite concentration of the negative controls and the residual nitrite of each microcosm after incubation. Linear models were used to assess the effects of the stress, *S*, the Gini coefficient, *G*, the relative abundance of the dominant species, *X*, and that of their interactions on ecosystem functionality. Both *G* and *X* are shape parameters describing a particular Lorenz curve. In addition to the factors considered above, several confounding factors could be present. Potential row, *R*, column, *C*, and experiment, *P*, effects due to the multiwell analysis process, negative controls, *B*, and the identity of the dominant species, *I*, were taken into account to allow a correct estimation of the model parameters being studied. Model selection was performed using a series of linear models in which each of the effects and interactions were entered sequentially (Table 1). After including the confounding factors, terms that resulted in the largest decrease of the AIC were added to the model. On the basis of the AIC, model 8 was selected (coefficient of multiple correlation, $R^2 = 78.5\%$). A residual analysis showed that the model fit was adequate (Supplementary Fig. 4).

From an ecological perspective, the assessment of stress, the identity of the dominant species and the shape parameters *G* and *X* are of great importance. There was a very significant effect on ecological functioning due to stress (chi-squared test, $P < 0.001$) and the latter's interactions with other variables (chi-squared test, $P < 0.001$). Hence, the type of stress had a strong impact on the functionality and on the contribution of the other variables in the model. Moreover, there was a very significant interaction between the identity of the dominant species and the stress. Chi-squared tests indicated that dominant species identity had no effect in the control environment ($P = 0.22$) or in the temperature-stressed environment ($P = 0.10$). The parameter estimates and tests (Supplementary Fig. 5) both showed that temperature had a negative impact on functionality. By contrast, the identity of the dominant species was shown to be significant (chi-squared test, $P < 0.01$) in the case of salt stress. This type of stress can therefore be considered selective, that is, one that disfavours some species but favours others (Supplementary Fig. 5). It should be noted that the functionality of none of the species was completely inhibited by temperature or salt stress (Supplementary Fig. 5).

The effect of the initial evenness on functionality and functional stability was modelled as a quadratic effect of the Gini coefficient (Fig. 2). The Gini coefficient was seen to have a very significant effect in both the control case and the salt-stress case ($P < 0.001$ for both tests). Both graphs (Figs 2a, c) show that functionality decreased with increasing initial unevenness and that this effect was more pronounced in the case of salt stress ($P < 0.001$). However, the adverse effect of initial unevenness can be partly overcome when the most dominant species is stress resistant, as illustrated by the interactions between the identity of the dominant species and salt stress. With regard to temperature, the degree of initial evenness had no significant effect, as growth was limited at low temperatures. Thus, in this situation, low temperature can be considered a severe, non-selective stress condition.

The type of stress had a distinct effect on the stress-buffering capability of a community (Fig. 3). A stress that disfavours all species to nearly the same extent decreases the functionality of the community regardless of its initial evenness. However, the degree of evenness is a key feature in cases of selective stress, which are the most frequent situations in nature[5,21,22]. We found that, on average, initial community unevenness decreases the functional stability when selective stress is applied. Nevertheless, exceptions occur, such as when the dominant species of an uneven community is favoured by the stress. Notably, increased initial community unevenness also lowered the functionality of unstressed communities, albeit not to the same extent as under selective stress.

Past practical and theoretical constraints have limited the ability to relate patterns of microbial evenness with the processes that determine

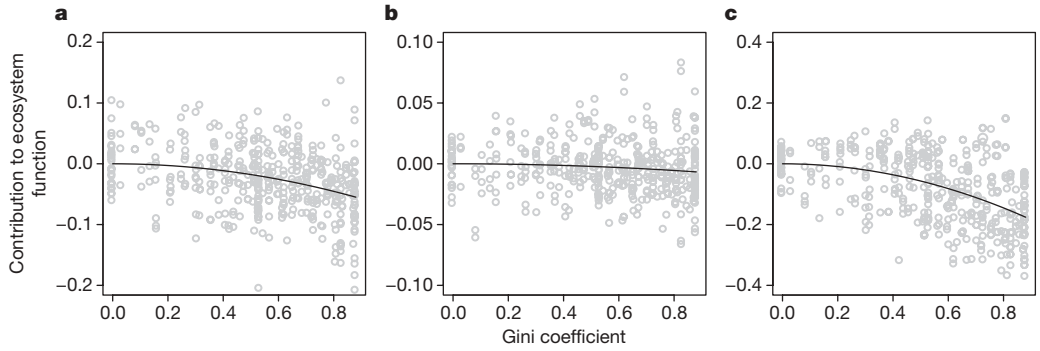


Figure 2 | Contribution of increasing initial unevenness (Gini coefficient) to the functionality of the ecosystem (that is, net denitrification after 20 h of incubation). **a**, No stress ($n = 420$); **b**, temperature stress ($n = 420$); **c**, salt stress ($n = 420$). This contribution to ecosystem function represents the effect of the Gini coefficient on the functionality corrected for row, column, experiment, negative control and main effect for stress. Partial residuals (contribution of the Gini coefficient plus residual) are indicated by open circles. They illustrate the extent of uncertainty that could not be explained by the model.

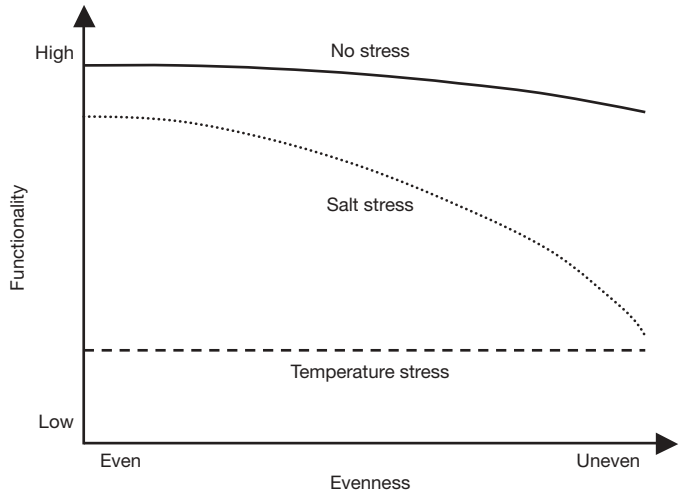


Figure 3 | Microbial functionality in relation to the initial evenness, for different types of stresses. The selective stressor NaCl has a much more negative impact on the functionality of the unevenly distributed microbial community.

these patterns. Nevertheless, recent studies have indicated that bacterial diversity may follow regular patterns, and that in some cases these patterns may be qualitatively similar to those observed for plants and animals[23]. Decreases in evenness (for example as a response to environmental changes) may have an indirect lowering effect on plant productivity[5]. Sparsely vegetated sites resulted in significantly lower evenness in bird communities[24]. Even in the field of palaeontology, it has already been postulated that the onset of less favourable environmental conditions is indicated by lower species evenness in arthropod and sponge communities[25].

Biodiversity protects ecosystems against declines in their functionality and allows for adaptation to changing conditions, because the coexistence of many species provides a greater guarantee that some will back up a given function when others fail[10,26]. Within the frame of this 'insurance hypothesis'[26], two aspects are important: (1) functional redundancy, in the sense of there being multiple species for each functional group[27,28], and (2) the relative abundances among these redundant species. At lower levels of species richness, the functionality of the ecosystem decreases[7]. In this research, all communities had the same degree of richness; hence, the importance of evenness for functional stability was isolated. Our results demonstrate that a community must have an even distribution among its functional redundant members if it is to respond rapidly to selective stress. In fact, when an ecosystem function in a highly uneven community depends strongly on the dominant species, the functional stability is endangered by environmental fluctuations[6]. Even under non-stressed conditions, high initial evenness is desirable for good functionality. Moreover, natural and anthropogenic activities influence the relative abundances more than the richness of species, and this has important consequences for ecosystems long before a species is threatened by extinction[5,6,21]. In conclusion, the existence of a highly diverse community, where redundant species may offer equivalent contributions to a specific function, may lead to higher functional stability during environmental fluctuations[6]. This implies that changes in community evenness should warrant increased attention in biodiversity surveys.

METHODS SUMMARY

Laboratory methods. The scheme of the experimental set-up is provided in Supplementary Fig. 1. A total of 18 denitrifying species were isolated from nature (Supplementary Table 3). Denitrifiers were classified by fatty-acid methyl ester analysis and 16S ribosomal RNA gene sequencing. The different operational taxonomic units were discriminated by repetitive extragenic palindromic PCR DNA fingerprinting. By analogy with ref. 7, we considered our operational taxonomic units as 'species'. Microcosms were obtained by mixing all 18 strains in different abundances. Nitrite was added to the mixtures that were incubated for 20 h without or with (temperature or salt) stress. The net ecosystem denitrification was estimated by the nitrite removal, which was measured spectrophotometrically (Sunrise, Tecan) as the difference of the absorbance at 540 nm before and after Montgomery reaction[29].

Experimental design and statistical analysis. Eighty-four different levels of initial evenness were possible, corresponding to a unique combination of Gini coefficient and X, the relative abundance of the dominant species, and each referred to as a design point (Supplementary Fig. 2). For the first experiments, each of the 84 design points was used twice. This resulted in 168 different microcosms that were placed on the multiwell plates *in duplo*. Additionally, 42 combinations of X and G were chosen according to an experimental design procedure to enable an optimal estimation of the linear and quadratic effects. The corresponding microcosms were placed *in duplo* on the multiwell plates. Model selection was performed using a series of linear models. Each of the variables and their interactions were entered sequentially and the models were compared on the basis of the AIC. The parameters of the mean model were estimated by ordinary least-squares methods. Following a residual analysis, the White estimator[30] was used to provide valid statistical inference in the presence of residuals with unequal variances (Supplementary Fig. 4).

Received 17 September 2008; accepted 28 January 2009.
Published online 8 March 2009.

1. Hooper, D. U. *et al.* Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. *Ecol. Monogr.* **75,** 3–35 (2005).
2. Loreau, M. *et al.* Biodiversity and ecosystem functioning: Current knowledge and future challenges. *Science* **294,** 804–808 (2001).
3. McCann, K. S. The diversity-stability debate. *Nature* **405,** 228–233 (2000).
4. Purvis, A. & Hector, A. Getting the measure of biodiversity. *Nature* **405,** 212–219 (2000).
5. Wilsey, B. J. & Potvin, C. Biodiversity and ecosystem functioning: Importance of species evenness in an old field. *Ecology* **81,** 887–892 (2000).
6. Balvanera, P., Kremen, C. & Martinez-Ramos, M. Applying community structure analysis to ecosystem function: examples from pollination and carbon storage. *Ecol. Appl.* **15,** 360–375 (2005).
7. Bell, T., Newman, J. A., Silverman, B. W., Turner, S. L. & Lilley, A. K. The contribution of species richness and composition to bacterial services. *Nature* **436,** 1157–1160 (2005).
8. Cardinale, B. J., Palmer, M. A. & Collins, S. L. Species diversity enhances ecosystem functioning through interspecific facilitation. *Nature* **415,** 426–429 (2002).
9. Loreau, M. & Hector, A. Partitioning selection and complementarity in biodiversity experiments. *Nature* **412,** 72–76 (2001).
10. Naeem, S. & Li, S. Biodiversity enhances ecosystem reliability. *Nature* **390,** 507–509 (1997).
11. Sankaran, M. & McNaughton, S. J. Determinants of biodiversity regulate compositional stability of communities. *Nature* **401,** 691–693 (1999).
12. Griffiths, B. S., Bonkowski, M., Roy, J. & Ritz, K. Functional stability, substrate utilisation and biological indicators of soils following environmental impacts. *Appl. Soil Ecol.* **16,** 49–61 (2001).
13. Huber, J. A. *et al.* Microbial population structures in the deep marine biosphere. *Science* **318,** 97–100 (2007).
14. Wilsey, B. J. &. P. o. l. l. e. y. H. W. Reductions in grassland species evenness increase dicot seedling invasion and spittle bug infestation. *Ecol. Lett.* **5,** 676–684 (2002).
15. Wu, T., Chellemi, D. O., Graham, J. H., Martin, K. J. & Rosskopf, E. N. Comparison of soil bacterial communities under diverse agricultural land management and crop production practices. *Microb. Ecol.* **55,** 293–310 (2008).
16. Yang, D. R., Peng, Y. Q., Yang, P. & Guan, J. M. The community structure of insects associated with figs at Xishuangbanna, China. *Symbiosis* **45,** 153–157 (2008).
17. Jessup, C. M. *et al.* Big questions, small worlds: microbial model systems in ecology. *Trends Ecol. Evol.* **19,** 189–197 (2004).
18. Kassen, R., Buckling, A., Bell, G. & Rainey, P. B. Diversity peaks at intermediate productivity in a laboratory microcosm. *Nature* **406,** 508–512 (2000).
19. Prosser, J. I. *et al.* The role of ecological theory in microbial ecology. *Nature Rev. Microbiol.* **5,** 384–392 (2007).
20. Philippot, L. & Hallin, S. Finding the missing link between diversity and activity using denitrifying bacteria as a model functional community. *Curr. Opin. Microbiol.* **8,** 234–239 (2005).
21. Chapin, F. S. III *et al.* Consequences of changing biodiversity. *Nature* **405,** 234–242 (2000).
22. Decho, A. W. Microbial biofilms in intertidal systems: an overview. *Cont. Shelf Res.* **20,** 1257–1273 (2000).
23. Horner-Devine, M. C., Carney, K. M. & Bohannan, B. J. M. An ecological perspective on bacterial biodiversity. *Proc. R. Soc. Lond. B* **271,** 113–122 (2004).
24. Symonds, M. R. E. & Johnson, C. N. Species richness and evenness in Australian birds. *Am. Nat.* **171,** 480–490 (2008).
25. Caron, J. B. & Jackson, D. A. Paleoecology of the Greater Phyllopod Bed community, Burgess Shale. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **258,** 222–256 (2008).

26. Yachi, S. & Loreau, M. Biodiversity and ecosystem productivity in a fluctuating environment: The insurance hypothesis. *Proc. Natl Acad. Sci. USA* **96,** 1463–1468 (1999).
27. Gitay, H., Wilson, J. B. & Lee, W. G. Species redundancy: A redundant concept? *J. Ecol.* **84,** 121–124 (1996).
28. Walker, B. H. Biodiversity and ecological redundancy. *Conserv. Biol.* **6,** 18–23 (1992).
29. Montgomery, H. A. C. & Dymock, J. F. The determination of nitrite in water. *Analyst* **86,** 414–416 (1961).
30. Kutner, M. H., Nachtsheim, C. J. & Neter, J. *Applied Linear Regression Models* 4th edn (McGraw-Hill Irwin, 2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to R. Amann for comments on the original manuscript and to P. Van Damme for practical assistance. This work was supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) (to L.W.), by an Interuniversity Attraction Pole research network grant of the Belgian government, Belgian Science Policy (to L.C.), by 'Program Master and Back' from Regione Sardegna (Italy; to A.B.), by 'Programma dell'Università per la Ricerca, PUR 2008' (ex FIRST) of the University of Milan (to D.D.), and by the Geconcerteerde Onderzoeksactie of Ghent University contract grant of the Ministerie van de Vlaamse Gemeenschap, Bestuur Wetenschappelijk Onderzoek (Belgium; to K.H., P.D.V., W.V. and N.B.).

Author Contributions L.W., M.M. and N.B. had the original idea for the experiment. The laboratory work was conducted by L.W., M.M., A.B. and K.H. The experimental design and statistical analyses were organized and performed by L.C. The manuscript was written principally by L.W., M.M. and L.C., with extensive input from D.D., K.H., P.D.V., W.V. and N.B.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.B. (nico.boon@ugent.be).

LETTERS

A micro-architecture for binocular disparity and ocular dominance in visual cortex

Prakash Kara[1] & Jamie D. Boyd[1]

In invertebrate predators such as the praying mantis and vertebrate predators such as wild cats the ability to detect small differences in inter-ocular retinal disparities is a critical means for accurately determining the depth of moving objects such as prey[1]. In mammals, the first neurons along the visual pathway that encode binocular disparities are found in the visual cortex. However, a precise functional architecture for binocular disparity has never been demonstrated in any species, and coarse maps for disparity have been found in only one primate species[2,3]. Moreover, the dominant approach for assaying the developmental plasticity of binocular cortical neurons used monocular tests of ocular dominance to infer binocular function[4]. The few studies that examined the relationship between ocular dominance and binocular disparity of individual cells used single-unit recordings and have provided conflicting results regarding whether ocular dominance can predict the selectivity or sensitivity to binocular disparity[5–9]. We used two-photon calcium imaging to sample the response to monocular and binocular visual stimuli from nearly every adjacent neuron in a small region of the cat visual cortex, area 18. Here we show that local circuits for ocular dominance always have smooth and graded transitions from one apparently monocular functional domain to an adjacent binocular region. Most unexpectedly, we discovered a new map in the cat visual cortex that had a precise functional micro-architecture for binocular disparity selectivity. At the level of single cells, ocular dominance was unrelated to binocular disparity selectivity or sensitivity. When the local maps for ocular dominance and binocular disparity both had measurable gradients at a given cortical site, the two gradient directions were orthogonal to each other. Together, these results indicate that, from the perspective of the spiking activity of individual neurons, ocular dominance cannot predict binocular disparity tuning. However, the precise local arrangement of ocular dominance and binocular disparity maps provide new clues regarding how monocular and binocular depth cues may be combined and decoded.

Binocular vision and depth discrimination evolved more than 100 million years ago[10]. In mammals, the first single-cell description of a binocular disparity detector in the brain was made in the cerebral cortex of the cat approximately 40 years ago[11]. Numerous single-unit studies followed in both cats and macaque monkeys, with pivotal electrophysiological and theoretical characterizations of the encoding of binocular disparity in the visual cortex, for example, position versus phase disparities and energy models[12–15]. In visual cortical neurons of mammals with frontally placed eyes, comparing the responses elicited by alternately stimulating each eye demonstrates the presence of a full range of ocular dominance, from completely contralateral to binocular to completely ipsilateral cells. A neuron tuned for binocular disparity, by definition, must receive visual input from both eyes. Therefore, it is reasonable to suppose that only binocular and not monocular cells would show robust disparity selectivity. A relationship is also suggested by misaligning the two eyes during the critical period of postnatal development. The misalignment leads to a loss of visual cortical neurons that can be driven through either eye[16] (that is, neurons lose their ocular dominance and become ocular exclusive—monocular). The misalignment also leads to stereo blindness[17]. Although ocular dominance is among the premier models of postnatal developmental plasticity[4], testing the input from each eye independently fails to show the suppressive effects or the summation of subthreshold inputs that can code for disparity in 'monocular' cortical cells[18]. Indeed, from single-unit electrophysiological studies, no consensus could be reached on the relationship between ocular dominance and binocular disparity[5–9].

By assaying binocular disparity and ocular dominance for nearly every neuron in a local volume of cat visual cortex using calcium imaging, we examined whether there is an orderly representation of binocular disparity and ocular dominance, and whether these two features were inter-related at the level of single cells and the local map structure. Our calcium indicator loading protocol typically labelled several hundred adjacent cortical layer 2/3 neurons in a spherical region (diameter 300–600 μm). Two-photon calcium imaging then permitted the simultaneous measurement of the visual responses of 100–150 neurons within a single optical cross-section parallel to the cortical surface. All 300 × 300 μm imaged sites were iso-orientation and iso-direction selective (Supplementary Figs 1a–c and 5b). Because we had to perform a battery of tests for ocular dominance, disparity, orientation, direction, spatial frequency and retinotopy, we did not probe disparity at orientation pinwheel sites. For our ocular dominance and binocular disparity measurements, we interleaved monocular and binocular drifting sine grating visual stimuli (see Fig. 1a). The two monocular stimuli and eight inter-ocular spatial phase disparity stimuli were always presented at the orientation and direction optimal for the imaged site (see Supplementary Figs 1a–c and 5b). Twelve cats were used in this study. In the first three animals, only ocular dominance was assessed ($n = 857$ cells) to determine the long-term stability of monocular responses over time (Supplementary Fig. 1d, e). In seven subsequent animals, monocular stimuli were always interleaved with binocular disparity stimuli ($n = 2{,}028$ cells). In two additional animals, imaging with simultaneous electrophysiological controls was performed (Supplementary Fig. 2).

With calcium imaging, individual visual cortical neurons showed robust and highly reproducible trial-by-trial responses to monocular stimuli and binocular disparity stimuli. Figure 1b shows the time course of the calcium indicator fluorescence signal evoked by visual stimulation for five simultaneously recorded cells from a single animal. Cell 1 had near equal responses to either monocular visual stimulus, robust responses to five of the eight presented disparity stimuli, and a clear suppression of visual responses to at least three binocular disparity stimuli (45°, 90° and 135° inter-ocular spatial phase disparities). Cell 3 responded almost exclusively to stimulation

[1]Department of Neurosciences, Medical University of South Carolina, Charleston, South Carolina 29425, USA.

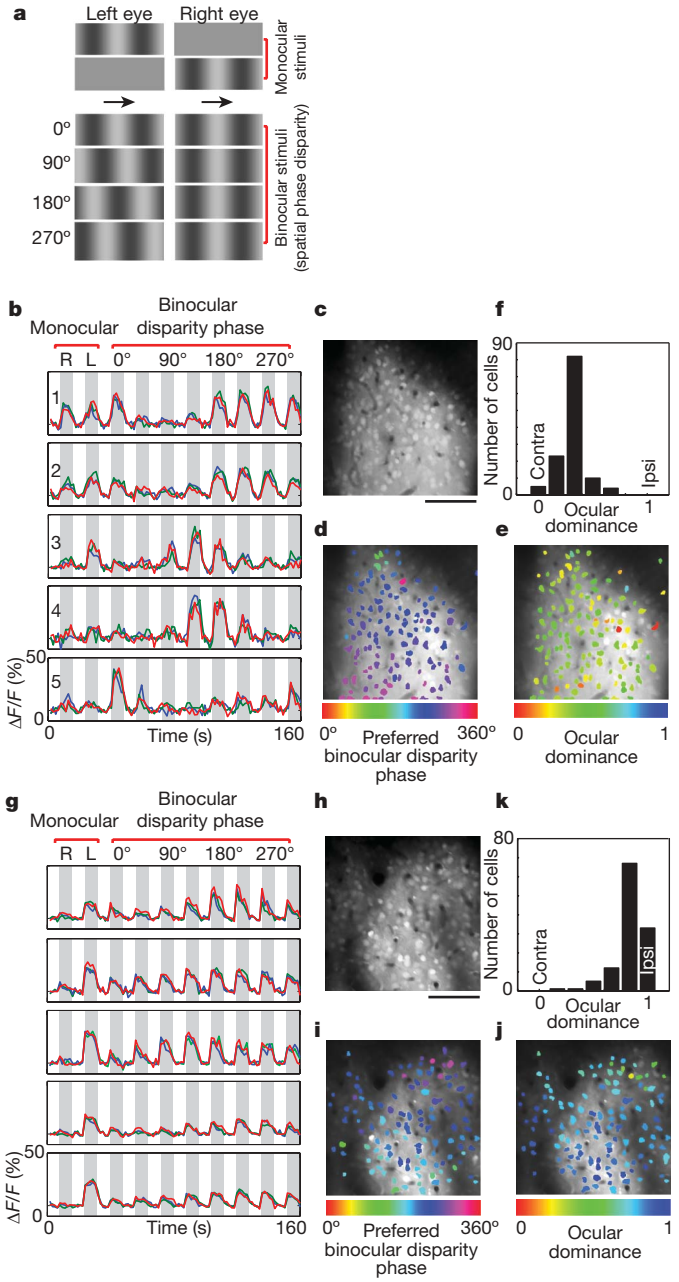


Figure 1 | Single-cell responses and functional maps from two experiments. **a**, Monocular and binocular stimuli used to obtain maps for ocular dominance and binocular disparity, respectively. Arrows pointing in the same direction denote that the grating stimuli presented to each eye always moved in the same direction during monocular and binocular viewing conditions. **b**, Time courses for five cells (numbered 1–5) from the site shown in **c**. Three trials are superimposed for each cell. Responses are shown for stimulation to the right eye (R), left eye (L) and then eight inter-ocular spatial phase binocular disparities in 45° steps. **c**, Calcium indicator loading in cells 201 μm below the pia. **d**, Cell-based binocular disparity map. Only cells significantly tuned for disparity are coloured: 119 out of 140 cells, $P < 0.05$, ANOVA across eight disparities. **e**, Cell-based ocular dominance map. **f**, Ocular dominance histogram. **g–k**, Data from a second animal for which monocular stimuli evoked responses primarily from the ipsilateral eye, but 114 of 124 cells were selectivity tuned for binocular disparity. The preferred orientation for cells at both cortical sites was 45° from vertical. Scale bars, 100 μm.

of the left eye when probed with monocular visual stimuli, but showed a profound modulation to binocular disparity stimulation with a peak response at 135° inter-ocular spatial phase disparity. Cell 4 had weak responses to monocular stimuli but again displayed potent modulation to specific phases of binocular disparity stimuli. Cell 5 was narrowly tuned to respond to a binocular spatial phase disparity of 0°.

The diversity of disparity tuning across neighbouring cells from a single 300 × 300 μm imaged site (for example, Fig. 1b) might imply the lack of a locally organized map for disparity. However, the cell-based disparity map (Fig. 1d) showed a smooth progression of preferred disparity phase from the bottom to the top of the imaged site. Only cells significantly tuned (selective) for binocular disparity phase were colour-coded in Fig. 1d and all subsequent cell-based disparity maps. For the data shown in Fig. 1d, of the 140 cells identified, 96% were significantly responsive to disparity stimuli ($P < 0.05$, analysis of variance (ANOVA) across blank and eight disparity periods) and 85% were tuned to binocular disparity ($P < 0.05$, ANOVA across eight disparity periods). In all colour-coded disparity phase maps, 0° preferred phase did not necessarily correspond to 0° absolute disparity. As demonstrated in previous studies in anaesthetized cats and monkeys[18–21], varying relative spatial phase disparity with sine gratings provides robust indices of disparity selectivity and sensitivity (also see Supplementary Discussion). The map for ocular dominance (Fig. 1e) from the same site as shown in Fig. 1d was relatively pure with virtually all cells being binocular with a slight contralateral bias, as confirmed in the ocular dominance histogram (Fig. 1f). Data from another cat are shown in Fig. 1g–k. The responses from five individual cells to monocular stimuli and binocular disparity stimuli were once again very robust (Fig. 1g). However, at this site monocular stimulation evoked responses almost exclusively from one eye (ipsilateral). Nevertheless, binocular disparity stimuli evoked significant and selective modulation of responses. The disparity map corresponding to this second site also showed a smooth transition of preferred disparity across the imaged area (Fig. 1i). As expected from the time courses shown in Fig. 1g, the ocular dominance map and histogram showed a strong bias to ipsilateral eye stimulation (Fig. 1j, k).

The two experiments described in Fig. 1 each have regions with a very narrow range of ocular dominance preferences. The fact that a strong map for disparity phase is present under both conditions indicates that disparity phase may be insensitive to ocular dominance. In sixteen 300 × 300 μm imaged areas from eleven calcium indicator dye injection sites in seven animals, not a single site showed a significant correlation between the cells' preferred disparity phase and their ocular dominance ($R = 0.001$ to 0.181; $P = 0.07$ to 0.99 per imaged area). The monocularity index[9], which ignores the sign of ocular dominance (ipsi versus contra) and quantifies only the strength of eye dominance, also did not yield a significant correlation with the cells' preferred disparity phase ($R = 0.001$ to 0.204; $P = 0.06$ to 0.99 per imaged area).

The independence of preferred disparity phase from ocular dominance at the level of single cells is best demonstrated when an individual 300 × 300 μm imaged site contained cells that had the full range of ocular dominance indices (Fig. 2). From qualitative observation of the cell-based maps for preferred disparity phase and ocular dominance (Fig. 2a, b), they appeared to be orthogonally oriented. We quantified the relative gradient direction of these maps by first smoothing the raw pixel maps for preferred disparity and ocular dominance (Fig. 2d, e). Smooth pixel maps were also derived from cell-based maps (see Methods) and produced almost identical results. The relative gradient direction of the ocular dominance and disparity maps was even more apparent when the disparity and ocular dominance maps were overlaid as contour plots (Fig. 2f). The relative gradient direction of the two maps was quantified by calculating the pixel-by-pixel difference in gradient direction for the two maps (Fig. 2g, Supplementary Figs 3 and 4, and Methods). A histogram of the distribution of the gradient direction difference for the two maps shows a clear peak near 90°, confirming that the maps were near perfectly orthogonal (also see Supplementary Fig. 3). From the fitted curve (red, Fig. 2g), we first calculated the ratio of the peak to the baseline (VMratio, see Methods).

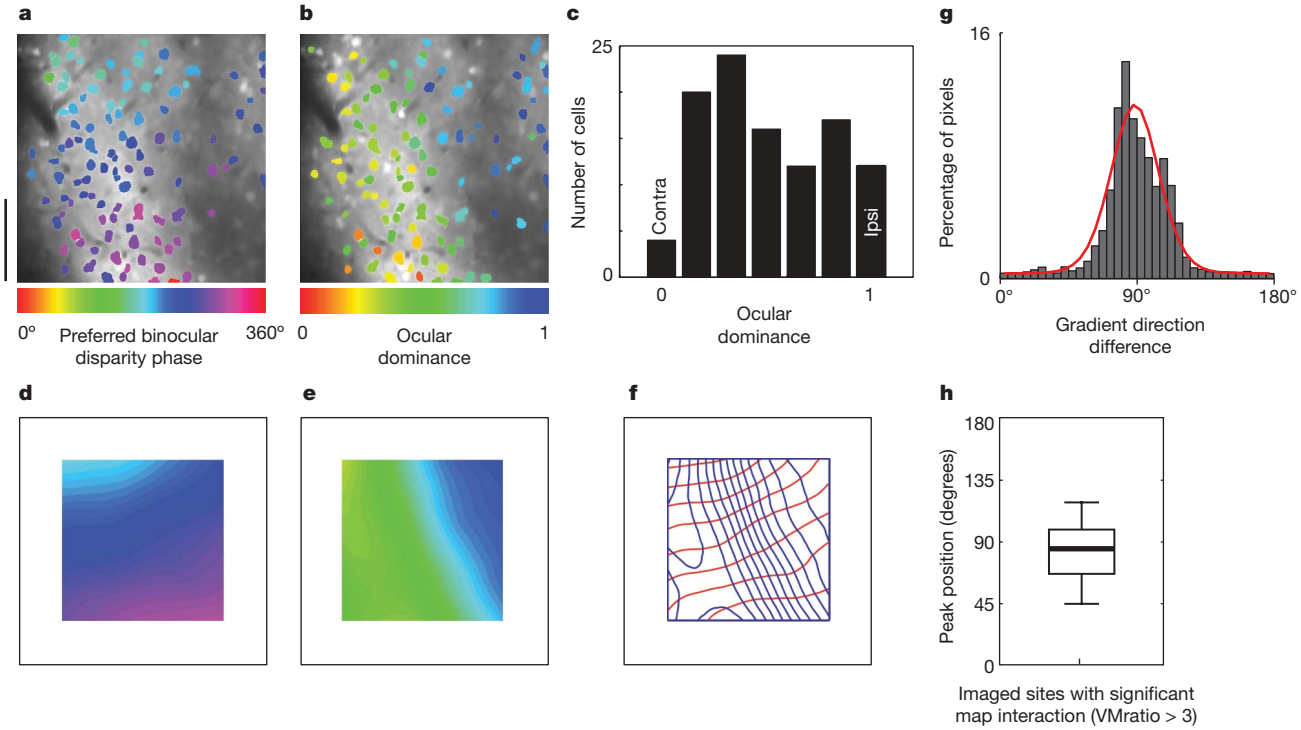


Figure 2 | Orthogonal maps for binocular disparity and ocular dominance when gradients were evident in both maps. **a**, **b**, Disparity and ocular dominance cell-based maps from a single imaged site 204 μm below the pia. **c**, Ocular dominance histogram shows that the complete range of ocular dominance indices (0–1) are represented at this site. The preferred orientation for cells at this site was vertical. **d**, **e**, Smoothed disparity (**d**) and ocular dominance (**e**) pixel-based maps used in the calculation for the difference in gradient direction for the two maps (**g**). To avoid artefacts, edges (52 μm on each side) were excluded from the analysis (also see Supplementary Fig. 9). **f**, Overlay of smoothed disparity and ocular dominance maps, each represented as contour plots (red for disparity, blue for ocular dominance), is indicative of orthogonality. **g**, Histogram of gradient direction difference for all pixels in the two maps show a peak centred near 90°. Red trace shows curve fitted to the histogram, peak at 88°, confirming orthogonality. **h**, Range of orthogonality for 8 out of 16 imaged areas (300 × 300 μm each) that had significant interaction. Bold horizontal line represents the median, boxes the 25% and 75% quartiles, and whiskers the 1% and 99% quantiles. Scale bar, 100 μm.

For each imaged site, a VMratio > 3 was considered to represent a significant interaction of the two maps ($n = 8$ out of 16 imaged areas, each 300 × 300 μm, showed significant interaction). The other eight imaged areas had no significant interaction (VMratio < 2). The lack of map interaction was further confirmed by randomizing the pixels in one of the two maps and showing that the gradient direction difference histograms still had a VMratio of less than 2. In the imaged areas (300 × 300 μm) where VMratio was less than 2, the gradient direction difference histograms from randomized versus non-randomized maps were statistically indistinguishable ($P = 0.404$ to 0.867, $Z = 0.16$ to 0.83, sign test). Thus, if the VMratio was less than 2, no interaction can be determined between two maps. Several examples of imaged sites that had either significant or no interaction are shown in Supplementary Fig. 4. The VMratio was correlated with the ocular dominance variance of cells (σ^2 cells) per imaged site ($R = 0.65$; $P < 0.01$; $n = 16$ imaged areas). Thus, significant interaction of the disparity phase and ocular dominance maps was more likely when the full range of ocular dominance was represented at a given site (for example, Fig. 2) compared to when a narrow range of ocular dominance was represented per site (for example, both cases in Fig. 1). We define 'orthogonality' as an angle difference in the range between 45° and 135°. For the eight 300 × 300 μm imaged areas that showed a significant ocular dominance versus disparity phase gradient interaction, the peak relative direction of the two gradients had a median angle of 85° with 25% and 75% quartile ranges falling within 66–99° (Fig. 2h, peak calculated from red curve fitted to histogram). Additional statistics on map interaction, for example, median scalar product, are given in the Supplementary Discussion.

Having shown that preferred binocular disparity phase is not related to ocular dominance at the level of single cells, we tested the possibility that ocular dominance might predict the sensitivity to binocular disparity. From qualitative observations, sites dominated by responses to monocular stimulation of either eye and sites responsive to monocular stimulation of only one eye appear to be just as likely to show robust disparity tuning. In the three imaged sites shown in Figs 1 and 2, 80–99% of cells were responsive and 80–92% of cells were selective for binocular disparity. Across all experiments in seven animals, 75% (1,512 out of 2,028) of cells were responsive to disparity stimuli. Of these visually responsive cells, 73% (1,097 out of 1,512) were tuned for binocular disparity. Thus, some individual 300 × 300 μm imaged sites only had ~30% of cells tuned for disparity. Furthermore, only 5% of all cells (101 out of 2,028) were truly monocular, that is, responsive to stimulation of one eye and not significantly responsive to disparity stimulation. To determine explicitly whether ocular dominance influenced the sensitivity to binocular disparity stimuli across our entire sample, we only considered cells that were significantly responsive to binocular disparity and monocular stimuli ($n = 1{,}119$ cells). Disparity sensitivity is reflected in the entire tuning curve for disparity, including facilitation and suppression relative to the mean response (Fig. 3a). We established that the ratio of the amplitude of a sine-fitted tuning curve for disparity to the mean response ($F1/F0$) was a reliable index of disparity sensitivity (Supplementary Fig. 6) and found that disparity sensitivity and ocular dominance were uncorrelated (Fig. 3b, $R = 0.041$, $P = 0.170$). $F1/F0$ was larger in sites where a significant map interaction between disparity and ocular dominance was measured (VMratio > 3), compared to sites where no map interaction was detected (VMratio < 2), that is, $F1/F0 = 0.653 \pm 0.011$ versus 0.485 ± 0.009, $P < 0.00001$, t-test). However, an $F1/F0$ of ~0.5 still represents very potent modulation.

Binocular disparity maps from a single injection site were stable over protracted time periods, up to the 12-h maximum time we

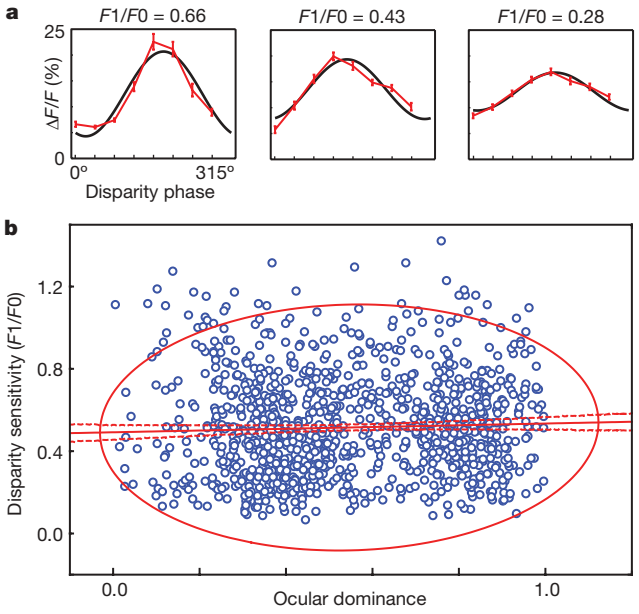


Figure 3 | Relationship between disparity sensitivity and the response to monocular stimuli. a, Disparity tuning curve for three cells. Data are shown in red, mean ± s.e.m., for the eight disparities presented. Sine fits are shown in black. The ratio of the amplitude of the sine fit to the measured mean response of the data ($F1/F0$) is a reliable measure of sensitivity to disparity (see Supplementary Fig. 6). **b**, Disparity sensitivity was uncorrelated with ocular dominance ($R = 0.041$, $P = 0.170$, $n = 1{,}119$ cells). Dashed lines are 95% confidence limits for the linear regression and the ellipse is the 95% prediction interval.

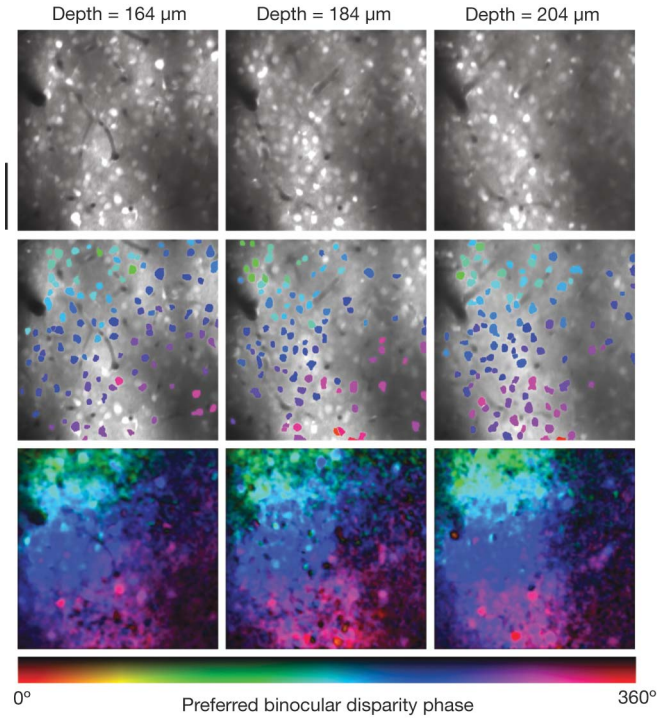


Figure 4 | Stable functional micro-architecture for binocular disparity. Anatomical images 300 × 300 μm (top row), cell-based disparity phase maps (middle row) and pixel-based disparity phase maps (bottom row) obtained at three depths (164, 184 and 204 μm) from a single site. Each data set was collected 60–90 min apart. The disparity gradient (degrees per μm) was similar for all three maps (mean ± s.d. = 0.518 ± 0.154; 0.585 ± 0.137; 0.514 ± 0.103). The preferred orientation for cells at this site was vertical. An additional data set from this site was collected nearly 12 h after the first (see Supplementary Fig. 3). Scale bar, 100 μm.

recorded from some sites (Fig. 4). The smoothness of the transition from one disparity domain to the next can best be appreciated from pixel maps (bottom row in Fig. 4). In these pixel maps, cell boundaries were ignored; the hue of each pixel was determined by the best disparity; and the brightness of each pixel was determined by the magnitude of the response vector. Such pixel maps therefore represent the combined response from cell bodies and surrounding neuropil[22]. The preferred disparity at all three depths changed systematically from the bottom right to the top left of each area and the mean disparity gradient was indistinguishable for the three depths (Fig. 4).

The trial-by-trial stability of responses reflected in the time course of fluorescence changes to disparity stimuli, the stable monocular retinotopy measurements (Supplementary Fig. 5) and the stability of the disparity maps over 12 h of recording indicate that we had no artefacts from eye movements or drift in our anaesthetized and paralysed animals. Small spatial frequency gradients were occasionally present in our imaged sites. However, they were orthogonal to the binocular disparity gradient (Supplementary Fig. 3). Imaging and electrophysiological controls indicated that changes in calcium fluorescence were not saturating and matched the spiking activity in individual cells (Supplementary Fig. 2). Additional analytical controls confirmed that any potential response onset transients did not confound our disparity tuning measurements at the level of single cells and the overall map structure (Supplementary Fig. 7).

Because we simultaneously recorded from at least 100 cells for each given site in cortical layer 2/3 with no sampling bias, it is unlikely that we missed an otherwise true correlation between ocular dominance and preferred binocular disparity phase or binocular disparity sensitivity. Perhaps a correlation between ocular dominance and binocular disparity will be found for simple cells in the primary recipient zone of thalamic input (cortical layer 4). However, complex cells in cat layer 2/3 are more ideally suited for disparity detection than simple cells[23], and most of our cells in layer 2/3 were binocular when probed with disparity stimuli.

The existence of a map for binocular disparity in area 18 of the cat visual cortex revealed with two-photon calcium imaging indicates that disparity maps may be more common across species with frontally placed eyes than previously thought. Individual iso-disparity domains in macaque extra-striate areas V2 and MT can be relatively large (750–1,500 μm), resulting in readily detectable maps for disparity with microelectrode or intrinsic imaging techniques[2,3]. For ocular dominance maps, we did not observe fractures (or jumps) in the map from ispilateral- to contralateral-eye-dominated regions in any circumstances. Transitions from binocular to apparently 'monocular' ocular dominance domains were always smooth. This indicates that the apparently weaker map structure for ocular dominance seen with conventional optical imaging methods[24] does not result from local mixing of neurons that have different ocular dominance indices.

The most comprehensive single-unit study so far in primate V1 did show independence between binocular disparity and ocular dominance at the level of single cells[9]. However, our two-photon calcium imaging experiments crystallize the exact relationship in the cat visual cortex by showing that the independence of disparity and ocular dominance at the level of single cells does not arise from a local salt-and-pepper arrangement of maps for either disparity or ocular dominance. From a developmental standpoint, a map for ocular dominance may initially reflect residual imbalances in the density of inputs from each eye[25]. However, a smooth map for ocular dominance may serve as a scaffold for the formation of disparity maps. Neurons embedded in local cortical regions where preferred disparity is organized in a map may be more sensitive to binocular disparity compared to adjacent regions that are less well organized, as is evident in primate MT (ref. 2). A potential computational advantage of the relationship between disparity and ocular dominance maps for binocular visual processing is that a wide range of disparity

encoding is maintained independent of local changes in ocular dominance. Locally organized ocular dominance and binocular disparity maps might optimize the processing of multiple depth cues by maximizing the coverage of binocular disparity and occlusion cues from surfaces located at different depths for which the 'eye of origin' needs to be known[26]. Cortical neurons tuned to this combination of features respond vigorously to monocular stimulation of one eye only but still show modest disparity tuning, for example, see cells 18, 58, 66, 71, 86, 95 and 106 in Supplementary Fig. 6b. Future studies could determine whether locally organized maps for ocular dominance and disparity have a role in speeding up the decoding (or readout) of these combined cues by other visual cortical areas. Although local orthogonality is an emergent property when multiple overlapping functional maps are simulated in the general class of self-organizing or dimension reduction models[27–29], it remains to be determined whether ocular dominance and disparity maps conform to various predictions made from such models.

METHODS SUMMARY

Cats (postnatal days 36–49) were anaesthetized with isoflurane (1–2% in surgery, 0.5–1.0% during imaging)[22] and paralysed with vecuronium bromide[22]. A craniotomy was performed over area 18 of the visual cortex, the dura reflected, and the underlying cortex covered with agarose. Movement of the brain from respiratory and heart-beat pulsations were negligible (Supplementary Fig. 8). The cell-permeant calcium indicator Oregon Green 488 Bapta-1 AM (1 mM) was prepared[22,30] and co-loaded with 40 μM Alexa Fluor 594 into a glass patch pipette (2.5 μm diameter tip). Under continuous visual guidance, the pipette tip was advanced 200–250 μm below the cortical surface and the indicators were then pressure-ejected (5–10 psi). This particular method of loading produces minimal staining of glial cells (see ref. 22) but it is possible that some of the stained cells in the present study were not neuronal. Fluorescence was monitored with a custom-built microscope (Prairie Technologies) coupled with a Mai Tai XF (Newport Spectra-Physics) mode-locked Ti:sapphire laser (850 nm or 920 nm). Drifting sine-wave gratings (2 Hz, 50% contrast) were presented on a CRT (100 Hz refresh rate) in a variety of configurations for orientation, direction of motion, spatial frequency, ocularity (left or right eye, for ocular dominance), and eight inter-ocular spatial phase disparities (0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°). For ocular dominance and binocular disparity assays, animals viewed the monoptic and dichoptic visual stimuli through ultra-fast ferroelectric liquid crystal shutters (7 kHz switching time, 1,000:1 extinction contrast ratio, DisplayTech). Each stimulus period (8 s) was preceded by an equal blank period, repeated 3–8 times. Coarse retinotopic positions of monocular receptive fields were determined by using 5° wide flashing bars of light or strips of gratings at ten retinotopic positions. Two-photon images were analysed in Matlab (Mathworks), see Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 July; accepted 5 December 2008.
Published online 21 January 2009.

1. Rossel, S. Binocular stereopsis in an insect. *Nature* **302,** 821–822 (1983).
2. DeAngelis, G. C. & Newsome, W. T. Organization of disparity-selective neurons in macaque area MT. *J. Neurosci.* **19,** 1398–1415 (1999).
3. Chen, G., Lu, H. D. & Roe, A. W. A map for horizontal disparity in monkey V2. *Neuron* **58,** 442–450 (2008).
4. Hubel, D. H. & Wiesel, T. N. Early exploration of the visual cortex. *Neuron* **20,** 401–412 (1998).
5. Poggio, G. F. & Fischer, B. Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *J. Neurophysiol.* **40,** 1392–1405 (1977).
6. Ferster, D. A comparison of binocular depth mechanisms in areas 17 and 18 of the cat visual cortex. *J. Physiol. (Lond.)* **311,** 623–655 (1981).
7. Gardner, J. C. & Raiten, E. J. Ocular dominance and disparity-sensitivity: why there are cells in the visual cortex driven unequally by the two eyes. *Exp. Brain Res.* **64,** 505–514 (1986).
8. LeVay, S. & Voigt, T. Ocular dominance and disparity coding in cat visual cortex. *Vis. Neurosci.* **1,** 395–414 (1988).
9. Read, J. C. & Cumming, B. G. Ocular dominance predicts neither strength nor class of disparity selectivity with random-dot stimuli in primate V1. *J. Neurophysiol.* **91,** 1271–1281 (2004).
10. Parker, A. *In the Blink of an Eye: How Vision Sparked the Big Bang of Evolution* (Basic Books, 2003).
11. Barlow, H. B., Blakemore, C. & Pettigrew, J. D. The neural mechanism of binocular depth discrimination. *J. Physiol. (Lond.)* **193,** 327–342 (1967).
12. DeAngelis, G. C., Ohzawa, I. & Freeman, R. D. Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature* **352,** 156–159 (1991).
13. Anzai, A., Ohzawa, I. & Freeman, R. D. Neural mechanisms underlying binocular fusion and stereopsis: position vs. phase. *Proc. Natl Acad. Sci. USA* **94,** 5438–5443 (1997).
14. Prince, S. J., Pointon, A. D., Cumming, B. G. & Parker, A. J. Quantitative analysis of the responses of V1 neurons to horizontal disparity in dynamic random-dot stereograms. *J. Neurophysiol.* **87,** 191–208 (2002).
15. Haefner, R. M. & Cumming, B. G. Adaptation to natural binocular disparities in primate V1 explained by a generalized energy model. *Neuron* **57,** 147–158 (2008).
16. Hubel, D. H. & Wiesel, T. N. Binocular interaction in striate cortex of kittens reared with artificial squint. *J. Neurophysiol.* **28,** 1041–1059 (1965).
17. Mitchell, D. in *The Visual Neurosciences* (eds Chalupa, L. M & Werner, J. S.) 189–204 (MIT Press, 2004).
18. Ohzawa, I. & Freeman, R. D. The binocular organization of simple cells in the cat's visual cortex. *J. Neurophysiol.* **56,** 221–242 (1986).
19. Freeman, R. D. & Ohzawa, I. Development of binocular vision in the kitten's striate cortex. *J. Neurosci.* **12,** 4721–4736 (1992).
20. Chino, Y. M., Smith, E. L. III, Hatta, S. & Cheng, H. Postnatal development of binocular disparity sensitivity in neurons of the primate visual cortex. *J. Neurosci.* **17,** 296–307 (1997).
21. Maruko, I. *et al.* Postnatal development of disparity sensitivity in visual area 2 (V2) of macaque monkeys. *J. Neurophysiol.* **100,** 2486–2495 (2008).
22. Ohki, K., Chung, S., Ch'ng, Y. H., Kara, P. & Reid, R. C. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* **433,** 597–603 (2005).
23. Ohzawa, I., DeAngelis, G. C. & Freeman, R. D. Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science* **249,** 1037–1041 (1990).
24. Bonhoeffer, T., Kim, D. S., Malonek, D., Shoham, D. & Grinvald, A. Optical imaging of the layout of functional domains in area 17 and across the area 17/18 border in cat visual cortex. *Eur. J. Neurosci.* **7,** 1973–1988 (1995).
25. Ringach, D. L. On the origin of the functional architecture of the cortex. *PLoS ONE* **2,** e251 (2007).
26. Shimojo, S., Silverman, G. H. & Nakayama, K. An occlusion-related mechanism of depth perception based on motion and interocular sequence. *Nature* **333,** 265–268 (1988).
27. Obermayer, K., Blasdel, G. G. & Schulten, K. Statistical–mechanical analysis of self-organization and pattern formation during the development of visual maps. *Phys. Rev. A* **45,** 7568–7589 (1992).
28. Swindale, N. V., Shoham, D., Grinvald, A., Bonhoeffer, T. & Hübener, M. Visual cortex maps are optimized for uniform coverage. *Nature Neurosci.* **3,** 822–826 (2000).
29. Yu, H., Farley, B. J., Jin, D. Z. & Sur, M. The coordinated mapping of visual space and response features in visual cortex. *Neuron* **47,** 267–280 (2005).
30. Stosiek, C., Garaschuk, O., Holthoff, K. & Konnerth, A. *In vivo* two-photon calcium imaging of neuronal networks. *Proc. Natl Acad. Sci. USA* **100,** 7319–7324 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Cumming and N. Swindale for discussions. We thank B. Shi, Z. Shen, Z. Lu and J. Schnellmann for comments on the manuscript. This work was supported by grants from the NIH, Whitehall and Dana Foundations to P.K.

Author Contributions P.K. conceived the project, designed the experiments and set up the laboratory. P.K. and J.D.B. performed the experiments. P.K. analysed the data and wrote the paper. Both authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.K. (kara@musc.edu).

METHODS

Images were analysed using customized Matlab (Mathworks) software. Cells were identified through a series of morphological filters that defined the contours of cell bodies based on intensity, size and shape[22]. Time courses of individual cells were extracted by calculating mean pixel values within cell contours[22]. Visually responsive cells were defined by ANOVA across blank and n test visual stimuli ($P < 0.05$). Cells selective for particular stimuli were defined by ANOVA across n stimulus periods ($P < 0.05$).

Ocular dominance (OD) was derived from the responses to monocular stimulation and defined as:

$$\mathrm{OD} = \frac{R_{\mathrm{ipsi}}}{\left(R_{\mathrm{ipsi}} + R_{\mathrm{contra}}\right)}$$

where R_{ipsi} is the response to ipsilateral eye stimulation and R_{contra} is the response to contralateral eye stimulation.

Sensitivity to binocular disparity ($F1/F0$) was derived by vector averaging as follows:

$$\mathbf{V}_{\mathrm{avg}} = \sum_{i=1}^{n} \frac{\mathbf{V}_i}{n}$$

where $\mathbf{V}_i$ is a vector with direction equal to disparity phase, length equal to the corresponding cell's response amplitude, and n is the total number of disparity phases.

The average amplitude of the response to disparity stimulation was defined as:

$$A_{\mathrm{avg}} = \sum_{i=1}^{n} \frac{|\mathbf{V}_i|}{n}$$

and then $F0$ and $F1$ were calculated as:

$$F0 = A_{\mathrm{avg}}$$

$$F1 = 2\left|\mathbf{V}_{\mathrm{avg}}\right|$$

The direction of $\mathbf{V}_{\mathrm{avg}}$ provided the phase of the best response to disparity stimulation.

Because of the low trial-by-trial variability of our data coupled with the use of sinusoidal grating visual stimuli, $F1/F0$ was a reliable index of sensitivity to disparity, as confirmed with Monte-Carlo-derived estimates of standard deviation of fit parameters and coefficients of determination (R^2).

Each time we calculated an analytical fit of experimental data we conducted Monte Carlo simulations (128 trials) to estimate the error of these analytical fits. For each Monte Carlo trial, we randomly modified values assuming they had Gaussian distributions with standard error as calculated from the analysis of the experimental data. Analytical fits were done for each simulated data set and the mean was calculated for all Monte Carlo trials. In all cases, the Monte-Carlo-derived means were nearly identical to the original data fit and we used the Monte-Carlo-derived standard deviations as error estimates of the fitting procedure. If cells passed the experimental alpha criteria for disparity selectivity ($P < 0.05$, ANOVA), then the mean $F1/F0$ was at least twice larger than the standard deviation derived from the Monte Carlo simulations (Supplementary Fig. 6c). For monocular retinotopy experiments, the Monte-Carlo-derived standard deviations were used to determine which experiments had retinotopic measures that were sufficiently reliable to use as an index of vergence state (see Supplementary Discussion).

To quantify the relative gradient direction of two maps, for example, disparity phase and ocular dominance, we first calculated the pixel-by-pixel gradient of smoothed pixel maps (compare to cell-based maps, below). We used a built-in Matlab function where the gradient (∇F) of a function of two variables $F(x,y)$ was defined as:

$$\nabla F = \frac{\partial F}{\partial x}\mathbf{X} + \frac{\partial F}{\partial y}\mathbf{Y}$$

To capture the global relationship between the two maps that have cellular structure, it was necessary to smooth the maps with a filter that is larger than the distance between two cells. Thus, each map was first lowpass Gaussian filtered with a standard deviation of 50 pixels (30 μm for 300 × 300 μm imaged regions). To remove small filtering artefacts present at edges (see Supplementary Fig. 9), borders around each map were excluded (52 μm on each side of 300 × 300 μm imaged regions; 105 μm on each side of 600 × 600 μm imaged regions).

For ocular dominance maps, the Gaussian filter was applied directly to pixel values of the ocular dominance map. Because disparity is a circular variable, an alternative smoothing procedure was used for the disparity phase map. First, two separate component maps (sine and cosine) were generated from the disparity angle map. Each component map was smoothed by the Gaussian filter. Next, each of the two smoothed component maps was combined back to a single disparity angle map.

To conduct an equivalent gradient analysis on cell-based maps, we first transformed cell-based maps to pixel maps as follows: we derived a value for each pixel $P_{x,y}$ by interpolating corresponding values from all cells surrounding each pixel. The interpolation was a weighted mean, where each weight was calculated as a Gaussian function of the distance to each cell:

$$P_{x,y} = \frac{\sum_{\mathrm{cell}=1}^{n} P_{\mathrm{cell}} W(x_{\mathrm{cell}}, y_{\mathrm{cell}}, x, y)}{\sum_{\mathrm{cell}=1}^{n} W(x_{\mathrm{cell}}, y_{\mathrm{cell}}, x, y)}$$

where $P_{x,y}$ is the new pixel value (disparity phase, ocular dominance) at each x,y coordinate in the map; P_{cell} is the corresponding cell-based value (disparity phase, ocular dominance); and W is the Gaussian function.

To maintain consistency with the Gaussian lowpass filter we used to smooth raw pixel maps; the standard deviation of the Gaussian function for the pixel maps used here was 50 pixels, which corresponds to 30 μm (for 300 × 300 μm imaged areas). Once pixel maps were generated from cell-based maps, the procedures for smoothing were identical as described earlier for raw pixel-based maps.

The gradient direction difference for two maps was calculated using the built-in Matlab function ∇F, as described previously. Because we were only interested in the relative direction of two simultaneously recorded maps—for example, disparity phase and ocular dominance—the gradient direction difference was collapsed to a 0–180° range (Fig. 2g). Each histogram was 36 bins in length and each bin represented 5 degrees.

To quantify the gradient direction difference distribution, we conducted two independent analyses of these histograms. First, using a least-squares method, we fit a von Mises function to the histogram:

$$G = A_{\mathrm{min}} + A_1 \exp\left\{A_2\left(\cos\left[(\mathrm{Ddir} - \mathrm{Ddir}_0)\frac{\pi}{90}\right] - 1\right)\right\}$$

where A_{min} is the value of the smallest bin in the distribution, Ddir is the gradient direction difference, and A_1, A_2 and Ddir_0 are fitting parameters.

The ratio of the maximum to the minimum of the fitted function (VMratio) was the first metric we used to quantify strength of the interaction of the two maps:

$$\mathrm{VMratio} = \frac{G_{\mathrm{max}}}{G_{\mathrm{min}}}$$

The second metric of the gradient direction difference histogram was calculated as the ratio of the number of pixels in 9 bins around the peak bin (max bin ± 4) to the total number of analysed pixels:

$$\mathrm{Bin\ ratio} = \frac{\sum_{N_{\mathrm{max}}-4}^{N_{\mathrm{max}}+4} N_n}{\sum_{1}^{36} N_n}$$

where N_n is the value of bin number n.

Pooling data from all imaged sites, these two measures of the strength of the map interaction were correlated ($R = 0.89$; $P < 0.00001$).

Vol 458|2 April 2009|doi:10.1038/nature07832

LETTERS

Decoding reveals the contents of visual working memory in early visual areas

Stephenie A. Harrison[1] & Frank Tong[1]

Visual working memory provides an essential link between perception and higher cognitive functions, allowing for the active maintenance of information about stimuli no longer in view[1,2]. Research suggests that sustained activity in higher-order prefrontal, parietal, inferotemporal and lateral occipital areas supports visual maintenance[3–11], and may account for the limited capacity of working memory to hold up to 3–4 items[9–11]. Because higher-order areas lack the visual selectivity of early sensory areas, it has remained unclear how observers can remember specific visual features, such as the precise orientation of a grating, with minimal decay in performance over delays of many seconds[12]. One proposal is that sensory areas serve to maintain fine-tuned feature information[13], but early visual areas show little to no sustained activity over prolonged delays[14–16]. Here we show that orientations held in working memory can be decoded from activity patterns in the human visual cortex, even when overall levels of activity are low. Using functional magnetic resonance imaging and pattern classification methods, we found that activity patterns in visual areas V1–V4 could predict which of two oriented gratings was held in memory with mean accuracy levels upwards of 80%, even in participants whose activity fell to baseline levels after a prolonged delay. These orientation-selective activity patterns were sustained throughout the delay period, evident in individual visual areas, and similar to the responses evoked by unattended, task-irrelevant gratings. Our results demonstrate that early visual areas can retain specific information about visual features held in working memory, over periods of many seconds when no physical stimulus is present.

To investigate the role of early visual areas in working memory, we used functional magnetic resonance imaging (fMRI) to monitor cortical activity while participants performed a delayed orientation discrimination task. During each trial, observers maintained fixation while two sample orientation gratings (~25° and ~115°) were briefly presented in randomized order, followed by a numerical cue indicating whether to remember the first or second grating (Fig. 1a). After an 11-s retention interval, a test grating was presented, and participants indicated which way it was rotated relative to the cued grating (±3° or ± 6°). This experimental design allowed us to isolate memory-specific activity. By presenting the same two gratings on every trial, we ensured that stimulus-driven activity could not predict the orientation held in working memory. It was also critical that the memory cue appeared after the presentation of the gratings and not beforehand. Otherwise, subjects could attend more to the appearance of the cued grating, which would enhance orientation-selective responses to that stimulus[17].

Behavioural data confirmed that observers could discriminate small differences in orientation between the cued grating and the test grating. Observers showed equally good performance when the first or second grating had to be remembered (75% and 73% correct, respectively, $T(5) = 1.24$, $P = 0.27$).

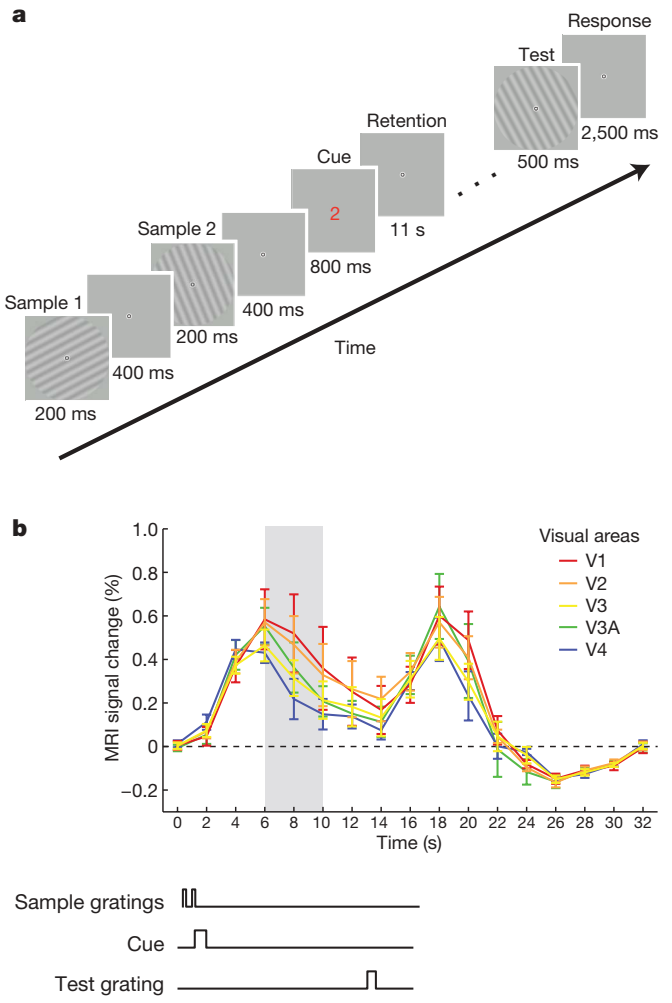


Figure 1 | Design of working memory experiment and resulting time course of fMRI activity. a, Timing of events for an example working memory trial. Two near-orthogonal gratings (25° ± 3°, 115° ± 3°) were briefly presented in randomized order, followed by a numerical cue (green '1' or red '2') indicating which grating to remember. After an 11-s retention period, a test grating was presented, and subjects reported whether it was rotated clockwise or anticlockwise relative to the cued grating. **b**, The time course of mean BOLD activity ($n = 6$) in corresponding regions of areas V1–V4 during the working memory task (0–16 s) and subsequent fixation period (16–32 s). Error bars indicate ± s.e.m. Time points 6–10 s (shaded grey area) were averaged for subsequent decoding analysis of delay-period activity. The start of this time window was chosen to allow for peak BOLD activity to fully emerge; we selected a conservative end point of 10 s to exclude any potential activity elicited by the test grating.

[1]Psychology Department and Vanderbilt Vision Research Center, Vanderbilt University, Nashville, Tennessee 37240, USA.

We used fMRI decoding methods to determine whether activity in early visual areas might reflect the contents of working memory (see Methods and Supplementary Methods). Although orientation selectivity primarily resides at fine spatial scales in the visual cortex, we have previously shown that pattern classification methods can successfully recover orientation information from cortical activity sampled at coarser resolutions using fMRI[17]. Here we investigated whether activity patterns during the delay period might predict which of the two orientations was held in working memory. For each trial, we calculated the average response of individual voxels over time points 6–10 s (Fig. 1b, grey region), selecting voxels from regions corresponding to 1–4° eccentricity in areas V1 to V4. The activity patterns observed on each trial served as input to a linear classifier with the cued orientation indicating the corresponding label. Classification accuracy was determined using cross-validation methods.

Ensemble activity pooled from areas V1–V4 was highly predictive of the orientation held in working memory, with prediction accuracy reaching 83% (Fig. 2, green curve). Decoding accuracy greatly exceeded chance-level performance of 50% ($T(5) = 18.2$, $P < 10^{-5}$), and proved highly reliable in each of the six participants (performance exceeding 58.75%, $P < 0.05$, binomial test). Notably, decoding was just as effective when the first grating was cued instead of the second (82.1% versus 83.6%, respectively, $T(5) = 1.0$, $P = 0.36$), indicating that this orientation information in the visual cortex was robust to potential interference from a subsequent item. Such robustness to interference has previously been found only in the prefrontal cortex[5]. Individual visual areas showed similar levels of orientation decoding performance ($F(3,15) = 1.71$, $P = 0.21$) ranging from 71–74% accuracy, with every participant showing above-chance decoding in each area. This indicates that maintaining an orientation in working memory is associated with widespread changes in orientation-selective activity throughout the early visual system, including V1, the first stage of orientation processing.

How do these orientation-selective responses for remembered gratings compare with stimulus-driven activity elicited by direct viewing of actual gratings? In a second experiment, participants had to identify letters presented rapidly at fixation while ignoring low-contrast oriented gratings (25° or 115°) flashing in the surround. Although the gratings were quite faint and task-irrelevant, they nonetheless evoked strong orientation-selective responses in early visual areas (Fig. 2, red curve). Activity in individual areas, V1, V2 and V3, was highly predictive of the orientation of the unattended gratings. Performance was considerably worse for V3A–V4 ($F(3,15) = 20.4$, $P < 10^{-4}$), presumably because activity in higher extrastriate areas is more dependent on visual attention[18]. Next, we evaluated the similarity of orientation-selective activity patterns in the two experiments by training the classifier on one data set and testing it on the other. Generalization performance for activity pooled across V1–V4 was below the performance found in the working memory experiment (Fig. 2, black curve), but was still significantly above chance ($T(5) = 6.0$, $P < 0.005$). Generalization was also better in V1 and V2 than in higher areas ($F(3,15) = 4.5$, $P < 0.05$), perhaps because these early areas exhibit stronger orientation-selective responses under stimulus-driven conditions[17]. Successful generalization across the two experiments is notable given how they differed in both stimulus and task. It seems that retaining an orientation in working memory recruits many of the same orientation-selective subpopulations as those that are activated under stimulus-driven conditions.

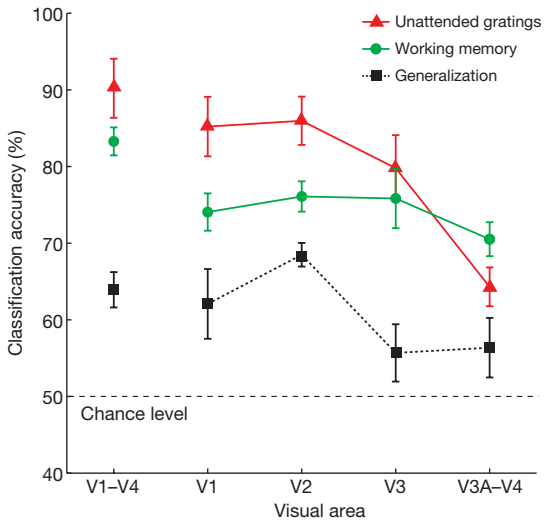


Figure 2 | Orientation decoding results for areas V1–V4. The accuracy of orientation decoding for remembered gratings in the working memory experiment (green circles), unattended presentations of low-contrast gratings (red triangles), and generalization performance across the two experiments (black squares). Error bars indicate ± s.e.m. Decoding was applied to the 120 most visually responsive voxels in each of V1, V2, V3 and V3A–V4 (480 voxels for V1–V4 pooled), as determined by their responses to a localizer stimulus (1–4° eccentricity). Individual areas V3A and V4 showed similar decoding performance but had fewer available voxels, so these regions were combined.

Further analyses confirmed that successful orientation decoding could not be explained by global differences in response amplitudes to the two orientations, as decoding applied to the averaged response of each visual area led to chance-level performance (46–57% accuracy, Supplementary Fig. 1a). We also tested for potential effects of global radial bias[19], and found that decoding was significantly impaired by spatially averaging the response of neighbouring voxels corresponding to different radial segments of the visual field (Supplementary Fig. 1b). In contrast, local variations in orientation preference within each radial segment led to high decoding accuracy (Supplementary Fig. 1c), consistent with the notion that much of the orientation information extracted by the classifier resulted from local anisotropies in orientation preference[17] (Supplementary Fig. 2).

Next, we investigated whether orientation-selective activity is maintained throughout the working memory delay period, by performing our decoding analysis on individual fMRI time points. Although individual functional images show poorer signal to noise, we could still detect changes in orientation-selective activity over time in both experiments. Orientation decoding of stimulus-driven activity in areas V1–V4 rose above chance level within 4 s of stimulus onset ($T(5) = 4.13$, $P < 0.01$) and reached asymptotic levels by ~6 s (consistent with the slow time course of the blood-oxygen-level-dependent (BOLD) response); performance remained high as gratings continued to be shown throughout the 16-s stimulus block (Fig. 3a). In comparison, orientation-selective activity in the working memory experiment was delayed by ~2 s, rising significantly above baseline by 6 s ($T(5) = 4.36$, $P < 0.01$) and reaching a plateau by 8 s. This delayed onset is consistent with the fact that observers did not see the task-relevant cue until 1.2 s after the first grating appeared, and required more time to interpret the cue. More notable is the fact that orientation-selective activity persisted throughout the delay period, when no physical stimulus was present, up until presentation of the test grating at time 13 s. Decoding of individual areas led to lower levels of performance; however, a similar pattern of results was found, as is shown for V1 (Fig. 3b).

Interestingly, this maintenance of orientation-selective information throughout the delay period did not seem to depend on a sustained boost in overall BOLD activity. The time course of mean BOLD activity for each visual area revealed a transient response to the first two gratings and a subsequent response to the test grating, with some suggestion of sustained activity in the intervening period (Fig. 1b). However, the level of sustained activity varied widely across subjects. For example, in V1 half of our subjects showed greater than baseline activity late in the delay period, whereas half did not (Supplementary Fig. 3a, b). Nevertheless, orientation-decoding performance was equally good for the two groups (74% versus 75%) and was sustained

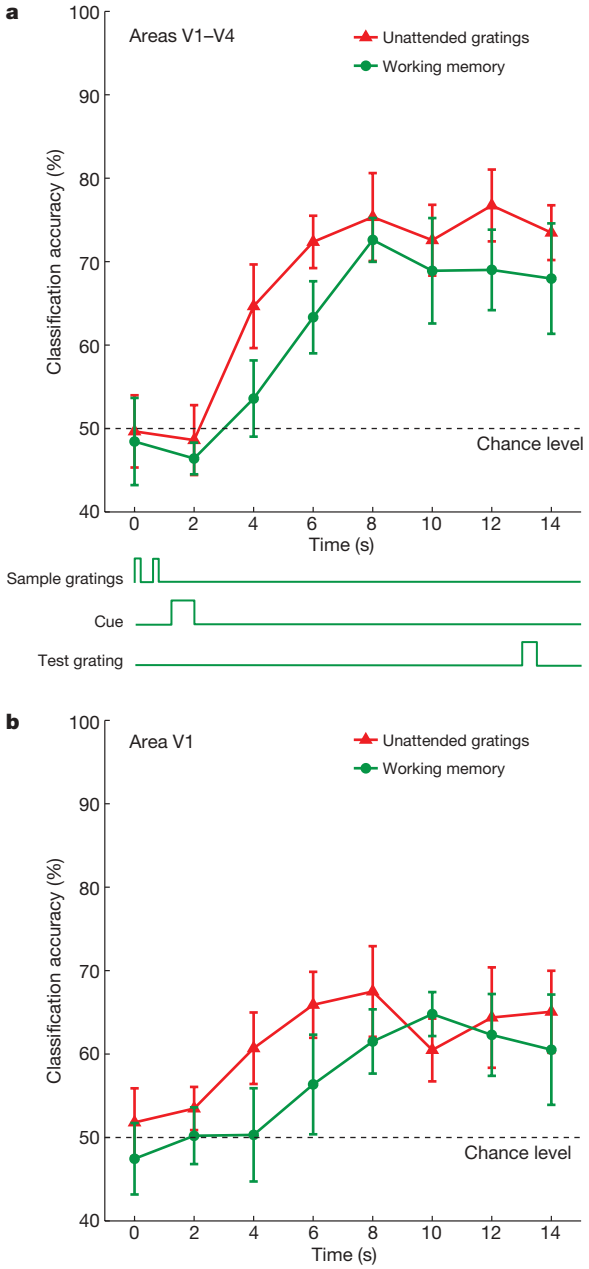


Figure 3 | Time-resolved decoding of individual fMRI time points. Orientation decoding of unattended stimulus gratings (red triangles), and remembered gratings during working memory (green circles), for activity obtained from areas V1–V4 (**a**) and from V1 only (**b**). Note that orientation information persists throughout the delay period during the working memory task, up until presentation of the test grating at time of 13 s. Error bars indicate ± s.e.m.

throughout the delay period (Supplementary Fig. 3c, d). Further analyses supported the notion that the overall BOLD amplitude from a region was unrelated to the amount of memory-related information available in the detailed activity pattern. We found no significant relationship between BOLD amplitudes and decoding accuracy across subjects, or across trials for individual subjects. Thus, it seems that low amplitude signals can nonetheless contain robust memory-related information throughout the entire delay period.

Additional control experiments indicated that this sustained orientation-selective activity reflected active maintenance of the cued orientation throughout the delay period rather than other cognitive processes. When observers were presented with a randomly selected pair of near-orthogonal orientations on every trial, it was still possible to decode which of the two orientations was held in working memory from activity in early visual areas (Supplementary Fig. 4). The use of randomly selected orientations ensured that long-term memory could not contribute to delayed discrimination; instead, accurate performance could only be achieved by maintaining the task-relevant grating seen on each trial (behavioural accuracy 76.2%). In another experiment, observers were shown two sample orientations followed by a numerical cue, the colour of which indicated whether to make a speeded judgment about the task-relevant orientation or to retain that orientation for subsequent discrimination. Whereas the immediate report task led to unreliable orientation decoding, active maintenance of the task-relevant grating over an extended 15-s delay led to sustained orientation-selective activity in areas V1 to V4 (Supplementary Fig. 5). Furthermore, we tested for effects of visual expectancy by omitting the sample gratings and providing only an initial cue to indicate the approximate orientation (~25° or ~115°) observers should expect at test. Expectation of a specific future orientation to be discriminated led to good behavioural performance (77.5% correct), but weak orientation-selective responses, as indicated by near chance-level decoding (Supplementary Fig. 6).

We also considered whether eye movements could account for successful decoding of remembered orientations; there are several reasons why this seems unlikely. First, sample gratings were presented for only 200 ms, too briefly for participants to prepare an eye movement within that time; also the working memory cue occurred afterwards, when no other stimulus was present. Second, an eye-tracking control experiment confirmed that all six participants maintained stable fixation when performing the working memory task (see Supplementary Methods). Unlike activity in the visual cortex, eye position signals failed to predict the orientation held in working memory (orientation decoding accuracy, 50.2%, $P = 0.94$). Third, it would be difficult to explain how strategic eye movements during working memory might elicit differential activity patterns that resemble those evoked by unattended gratings when participants had to attend to letters at fixation. Both the stimulus conditions and the strategic demands of the two experiments were profoundly different.

Our results provide new evidence to show that early visual areas can retain specific information about visual features held in working memory. When participants had to remember a precise orientation, this information was maintained in sensory areas, including the primary visual cortex where orientation tuning is strongest. Although V1 is essential for low-level feature processing, there is increasing evidence to suggest a role for V1 in conscious perception[20], attentional selection[18,20] and more complex cognitive functions[21,22]. We find that early visual areas are not only important for processing information about the immediate sensory environment, but can also maintain information in the absence of direct input to support higher-order cognitive functions.

Thus far, there has been little evidence to link V1 activity to visual working memory, perhaps because these tasks do not normally lead to increased activity in the visual cortex[14–16]. One study did find relatively greater V1 activity when monkeys had to report a remembered spatial location by means of an eye movement[23], but this increase in baseline activity could reflect the effects of spatial attention[18,24] or eye movement preparation[25]. Here we found that the overall activity in the visual cortex fell to near-baseline levels after prolonged delays, yet decoding of these low amplitude signals led to reliable prediction of the orientation held in memory.

Our findings suggest a potentially important source of memory-related information that may have been overlooked in previous studies, and indicate promising avenues for future research. Assuming that items in visual working memory are encoded by low levels of population activity, the application of population-decoding methods could help to uncover the underlying neural representations. Previous attempts to decode remembered information from delay-period activity in single neurons have typically led to low or chance levels of performance[5,16,26]. Perhaps if signals from many neurons or neuronal sites were recorded simultaneously to exclude the effects of correlated noise[27], far greater

information could be uncovered about items retained in memory, as was demonstrated here. The role of synaptic activity in the visual cortex might also be useful to explore, given that the BOLD response is more strongly associated with synaptic than spiking activity[28]. One recent study has reported suggestive evidence of enhanced local field potentials (4–10 Hz) in area V4 of the monkey during a visual working memory task[29]. Curiously, spiking activity did not increase overall but it was more likely to be observed at a specific phase of these slow oscillations, suggesting that the relationship between working memory and spiking activity might go beyond simple changes in firing rate.

It will be interesting for future studies to investigate whether working memory information found in the visual cortex is actively maintained by long-range recurrent interactions between higher-order areas and early visual areas, local recurrent activity within early visual areas, or a combination of both mechanisms. Presumably, prefrontal or parietal areas contributed to the top-down selection process, given that participants had to interpret an abstract cue indicating which of two orientations to hold in memory. However, it has been debated whether feedback signals from higher-order areas would necessarily reflect the contents of working memory[8]. Most network models of working memory have emphasized the importance of local recurrent activity[30]. In these models, a specific pattern of activity can be sustained after stimulus removal if units tuned to similar features share strong excitatory connections, balanced by broad inhibition from units tuned to other features. It is possible that the functional organization of orientation-selective neurons in the visual cortex could provide an infrastructure for such interactions. The present results demonstrate that early visual areas can indeed sustain information for periods of many seconds, indicating that their function is not restricted to sensory processing but extends to the maintenance of visual features and patterns in memory.

METHODS SUMMARY

Six observers, aged 24–36, with normal or corrected-to-normal vision, participated in this study, after providing written informed consent. The study was approved by the Vanderbilt University Institutional Review Board.

The main study consisted of three fMRI experiments. The working memory experiment involved delayed discrimination of one of two randomly cued orientations (Fig. 1a). Sine-wave gratings were centrally presented at ~25° or ~115° orientation (radius 5°, contrast 20%, spatial frequency 1 cycle per degree, randomized phase). The unattended gratings experiment required participants to report whenever a 'J' or 'K' appeared within a sequence of centrally presented letters (4 letters per s, performance accuracy 87.3%) while task-irrelevant gratings flashed on or off every 250 ms during each 16-s stimulus block. Gratings were identical to those used in the working memory experiment, but presented at lower contrast (4%) to elicit weaker visual responses, as might be expected during working memory. The visual-field localizer experiment consisted of blocked presentations of flickering random dots (dot size, 0.2°; display rate, 10 images per s), presented within an annulus of 1–4° eccentricity. This smaller window was used to minimize selection of retinotopic regions corresponding to the edges of the grating stimuli. Observers were instructed to maintain fixation on a central bull's eye throughout every experiment. Participants completed 8–10 working memory runs (32–40 trials per orientation), 4–5 unattended grating runs (28–35 blocks per orientation), and 2 visual-field localizer runs.

Scanning was performed using a 3.0-Tesla Philips Intera Achieva MRI scanner at the Vanderbilt University Institute of Imaging Science. We used gradient-echo echoplanar T2*-weighted imaging (time to echo (TE), 35 ms; repetition time (TR) 2,000 ms; flip angle, 80°; 28 slices, voxel size, 3 × 3 × 3 mm) to obtain functional images of the entire occipital lobe, as well as posterior parietal and temporal regions. Participants used a bite bar system to minimize head motion.

Received 7 August 2008; accepted 29 January 2009.
Published online 18 February 2009.

1. Baddeley, A. Working memory: looking back and looking forward. *Nature Rev. Neurosci.* **4,** 829–839 (2003).
2. Luck, S. J. & Vogel, E. K. The capacity of visual working memory for features and conjunctions. *Nature* **390,** 279–281 (1997).
3. Fuster, J. M. & Alexander, G. E. Neuron activity related to short-term memory. *Science* **173,** 652–654 (1971).
4. Miyashita, Y. & Chang, H. S. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* **331,** 68–70 (1988).
5. Miller, E. K., Erickson, C. A. & Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16,** 5154–5167 (1996).
6. Courtney, S. M., Ungerleider, L. G., Keil, K. & Haxby, J. V. Transient and sustained activity in a distributed neural system for human working memory. *Nature* **386,** 608–611 (1997).
7. Pessoa, L., Gutierrez, E., Bandettini, P. & Ungerleider, L. Neural correlates of visual working memory: fMRI amplitude predicts task performance. *Neuron* **35,** 975–987 (2002).
8. Curtis, C. E. & D'Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends Cogn. Sci.* **7,** 415–423 (2003).
9. Todd, J. J. & Marois, R. Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature* **428,** 751–754 (2004).
10. Vogel, E. K. & Machizawa, M. G. Neural activity predicts individual differences in visual working memory capacity. *Nature* **428,** 748–751 (2004).
11. Xu, Y. & Chun, M. M. Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature* **440,** 91–95 (2006).
12. Magnussen, S. & Greenlee, M. W. The psychophysics of perceptual memory. *Psychol. Res.* **62,** 81–92 (1999).
13. Pasternak, T. & Greenlee, M. W. Working memory in primate sensory systems. *Nature Rev. Neurosci.* **6,** 97–107 (2005).
14. Offen, S., Schluppeck, D. & Heeger, D. J. The role of early visual cortex in visual short-term memory and visual attention. *Vision Res.* doi:10.1016/j.visres.2007.12.022 (in the press).
15. Bisley, J. W., Zaksas, D., Droll, J. A. & Pasternak, T. Activity of neurons in cortical area MT during a memory for motion task. *J. Neurophysiol.* **91,** 286–300 (2004).
16. Zaksas, D. & Pasternak, T. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *J. Neurosci.* **26,** 11726–11742 (2006).
17. Kamitani, Y. & Tong, F. Decoding the visual and subjective contents of the human brain. *Nature Neurosci.* **8,** 679–685 (2005).
18. Kastner, S. & Ungerleider, L. G. Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci.* **23,** 315–341 (2000).
19. Sasaki, Y. *et al.* The radial bias: a different slant on visual orientation sensitivity in human and nonhuman primates. *Neuron* **51,** 661–670 (2006).
20. Tong, F. Primary visual cortex and visual awareness. *Nature Rev. Neurosci.* **4,** 219–229 (2003).
21. Kosslyn, S. M., Ganis, G. & Thompson, W. L. Neural foundations of imagery. *Nature Rev. Neurosci.* **2,** 635–642 (2001).
22. Roelfsema, P. R. Elemental operations in vision. *Trends Cogn. Sci.* **9,** 226–233 (2005).
23. Super, H., Spekreijse, H. & Lamme, V. A. A neural correlate of working memory in the monkey primary visual cortex. *Science* **293,** 120–124 (2001).
24. Ress, D., Backus, B. T. & Heeger, D. J. Activity in primary visual cortex predicts performance in a visual detection task. *Nature Neurosci.* **3,** 940–945 (2000).
25. Geng, J. J., Ruff, C. C. & Driver, J. Saccades to a remembered location elicit spatially specific activation in the human retinotopic visual cortex. *J. Cogn. Neurosci.* **21,** 230–245 (2009).
26. Miller, E. K., Li, L. & Desimone, R. Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J. Neurosci.* **13,** 1460–1478 (1993).
27. Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nature Rev. Neurosci.* **7,** 358–366 (2006).
28. Logothetis, N. K. *et al.* Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412,** 150–157 (2001).
29. Lee, H., Simpson, G. V., Logothetis, N. K. & Rainer, G. Phase locking of single neuron activity to theta oscillations during working memory in monkey extrastriate visual cortex. *Neuron* **45,** 147–156 (2005).
30. Wang, X. J. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* **24,** 455–463 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Brady and B. Wolfe for technical support, and J. Gore and the Vanderbilt University Institute of Imaging Science for MRI support. This work was supported by a grant from the National Eye Institute, National Institutes of Health to F.T. and a postgraduate fellowship from the Natural Sciences and Engineering Research Council of Canada to S.A.H.

Author Contributions F.T. devised and designed the experiments, S.A.H. and F.T. programmed the experiments, S.A.H. conducted the experiments and carried out the analyses with assistance from F.T., F.T. and S.A.H. wrote the paper together.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.T. (frank.tong@vanderbilt.edu).

Vol 458|2 April 2009|doi:10.1038/nature07930

LETTERS

Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals

Johannes F. Scheid[1,6], Hugo Mouquet[1], Niklas Feldhahn[1], Michael S. Seaman[7], Klara Velinzon[1], John Pietzsch[1,8], Rene G. Ott[2], Robert M. Anthony[2], Henry Zebroski[3], Arlene Hurley[4], Adhuna Phogat[9], Bimal Chakrabarti[9], Yuxing Li[9], Mark Connors[10], Florencia Pereyra[11], Bruce D. Walker[11], Hedda Wardemann[12], David Ho[13], Richard T. Wyatt[9], John R. Mascola[9], Jeffrey V. Ravetch[2] & Michel C. Nussenzweig[1,5]

Antibodies to conserved epitopes on the human immunodeficiency virus (HIV) surface protein gp140 can protect against infection in non-human primates, and some infected individuals show high titres of broadly neutralizing immunoglobulin (Ig)G antibodies in their serum. However, little is known about the specificity and activity of these antibodies[1–3]. To characterize the memory antibody responses to HIV, we cloned 502 antibodies from HIV envelope-binding memory B cells from six HIV-infected patients with broadly neutralizing antibodies and low to intermediate viral loads. We show that in these patients, the B-cell memory response to gp140 is composed of up to 50 independent clones expressing high affinity neutralizing antibodies to the gp120 variable loops, the CD4-binding site, the co-receptor-binding site, and to a new neutralizing epitope that is in the same region of gp120 as the CD4-binding site. Thus, the IgG memory B-cell compartment in the selected group of patients with broad serum neutralizing activity to HIV is comprised of multiple clonal responses with neutralizing activity directed against several epitopes on gp120.

During HIV infection some patients develop high titres of broadly neutralizing antibodies. However, despite intensive study over two decades, only a small number of broadly neutralizing monoclonal antibodies have been identified. These antibodies can be protective against chimaeric simian immunodeficiency virus (SIV)/HIV infection in macaques, and exert selective pressure on the virus[4–7]. Therefore, it is widely believed that such antibodies may be important components of any vaccine[1–3]. We thus set out to understand the naturally occurring memory antibody response in HIV-infected individuals who developed high titres of broad neutralizing serological activity.

Artificially trimerized gp140 protein composed of gp120 and gp41 was used to purify HIV-specific memory B cells from the blood of six patients, and immunoglobulin heavy and light chains were cloned from single-cell complementary DNA libraries[8,9] (Fig. 1a, Supplementary Figs 1–3 and Supplementary Tables 1 and 2). In contrast to random antibody cloning from memory B cells[10,11], and to the antibodies isolated from B cells that did not bind to gp140 from the same subjects, we found many clonally related antibodies in the gp140-binding B cells (Fig. 1b, Supplementary Figs 3 and 4 and Supplementary Table 3). The number of B-cell clones varied among patients from 22 to 50, and each clone was differentially expanded (Fig. 1b and Supplementary Table 3). Individual IgGs were expressed by transfection and tested for reactivity by enzyme-linked immunosorbent assays (ELISA). Eighty-six per cent of the antibodies cloned from gp140-binding B cells were gp140 reactive (Fig. 1b). In contrast, none of the antibodies obtained from the non-gp140-binding cells was gp140 specific (Fig. 1d and Supplementary Table 3). Out of 502 antibodies, we obtained 433 that bound to gp140, comprising 134 different B-cell clones (Supplementary Table 3).

When compared to IgG antibodies derived from non-gp140-binding B cells or historical controls[11] the gp140-binding antibodies were enriched for heavy-chain variable region 1 (V_H1)[12], immunoglobulin light-chain kappa (Igκ) versus immunoglobulin light-chain lambda (Igλ), and joining segment kappa 2 (Jκ2) or Jκ5 (Figs 1c and 2a–c). Individual patients showed longer or more charged IgH complementarity-determining region 3s (CDR3s), but these features were not found in all patients (Fig. 2b and Supplementary Fig. 5). An unexpected finding was that anti-gp140 antibodies were highly mutated (Fig. 2d and Supplementary Fig. 6). We conclude that anti-gp140 memory B cells are strongly selected post-germinal centre cells skewed to Igκ and V_H1 use. The exceptionally high level of mutation found in these antibodies may reflect chronic immune responses to HIV and persistent hypermutation and selection.

To map the antigenic specificity of the gp140-binding antibodies we performed ELISA experiments with purified gp120 and gp41. Seventy per cent of the gp140 antibodies bound to gp120, and 30% bound to gp41 (Fig. 3a–c). None of the 132 anti-gp41 antibodies assayed bound to the membrane proximal peptides recognized by the two broadly neutralizing anti-gp41 monoclonal antibodies 2F5 and 4E10 (refs 13, 14), and only nine antibody clones bound to the reported immunodominant region of gp41 (ref. 15) (Supplementary Table 3). Thus, most of the gp41 antibodies in the patients studied recognize conformational determinants, and antibodies to the membrane proximal region are difficult to detect despite the fact that both 2F5 and 4E10 bind to the trimer, which also absorbs most of the anti-gp41 antibodies in the patients' serum (Fig. 3c and Supplementary Fig. 2).

The specificity of the anti-gp120-binding antibodies was mapped using a collection of mutant proteins: gp120(D368R) interferes with binding to CD4 and all known anti-CD4-binding site (hereafter termed anti-CD4bs) antibodies, including b12 (refs 16–18); gp120(I420R) interferes with CD4-induced co-receptor-binding site antibodies (anti-CD4i), including 17b (ref. 19); gp120 core lacks the variable loops (VLs) and interferes with anti-VL and CD4i antibodies[20]. Antibodies that bound to gp120, gp120 core and gp120(I420R), but not to gp120(D368R), were classified as CD4bs-directed. Similarly, those that

[1]Laboratory of Molecular Immunology, [2]Laboratory of Molecular Genetics and Immunology, [3]Proteomics Resource Center, [4]Rockefeller University Hospital, and [5]Howard Hughes Medical Institute, The Rockefeller University, New York, New York 10065, USA. [6]Charite Universitaetsmedizin, D-10117 Berlin, Germany. [7]Beth Israel Deaconess Medical Center, Boston, Massachusetts 02215, USA. [8]Institute of Chemistry and Biochemistry, Freie Universität Berlin, D-14195 Berlin, Germany. [9]Vaccine Research Center, and [10]Laboratory of Immunoregulation, National Institutes of Allergy and Infectious Diseases, National Institutes of Health Bethesda, Maryland 20892, USA. [11]Partners AIDS Research Center, Mass General Hospital and Harvard Medical School, Charlestown, Massachusetts 02129, USA. [12]Max Planck Institute for Infection Biology, D-10117 Berlin, Germany. [13]Aaron Diamond Aids Research Center; New York, New York 10065, USA.

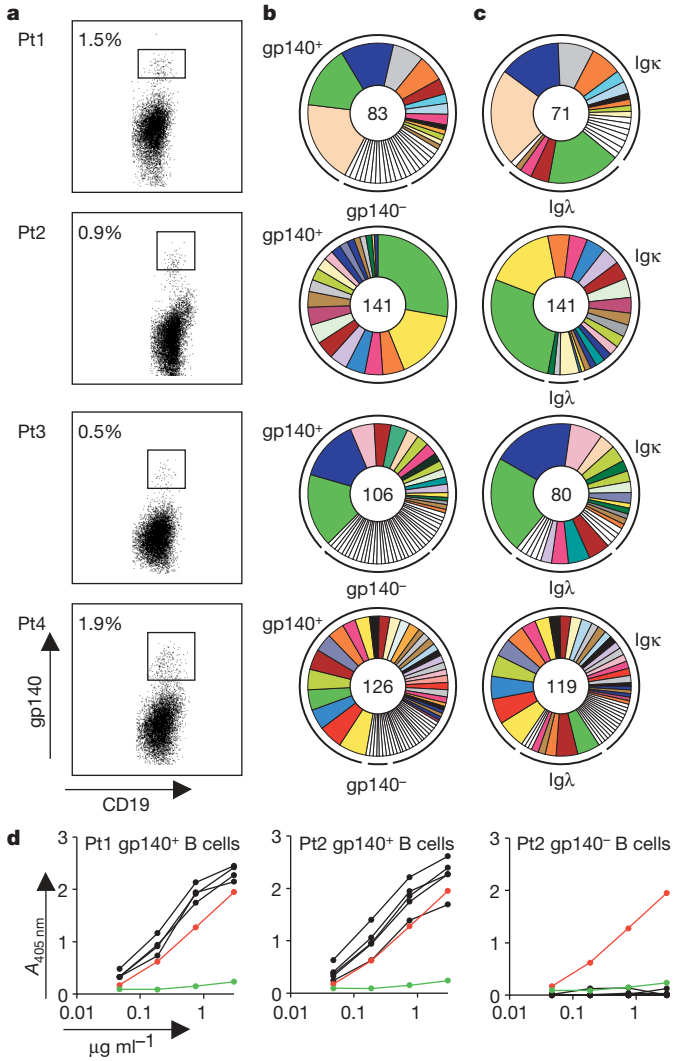


Figure 1 | Anti-gp140 antibody cloning. **a**, Flow cytometry plots of peripheral blood mononuclear cells from four HIV patients (Pt1–Pt4) stained with anti-CD19 and biotin–gp140. **b**, Distribution of gp140-binders ($gp140^+$) and non-binders ($gp140^-$) among all antibodies cloned. **c**, Igκ and Igλ expression among all gp140-binding antibodies. The number in the centre of the pies denotes the number of antibodies; slices are unique clones and proportional to clone size. **d**, gp140-binding ELISA for antibodies from gp140-binding cells (from patients 1 or 2) and from non-binders (from patient 2). The red line shows the b12 (ref. 25) control, the green line is negative control mGO53 (ref. 29).

bound to gp120 and gp120(I420R), but not to gp120 core, were classified as anti-VL antibodies, and those that bound to gp120 and gp120(D368R), but not to gp120(I420R), were classified as anti-CD4i antibodies. Anti-CD4bs, anti-CD4i and anti-VL antibodies were found in all four of the more complete patients but their relative representation varied markedly (Fig. 3b). Among all anti-gp140 antibodies, anti-CD4bs made up 9%, anti-CD4i 15% and anti-VL 27% (Fig. 3b). Only three of the anti-gp120 antibody clones bound to linear peptides and all of these to a region within the V3 loop[21] (Supplementary Table 3)

Surface plasmon resonance experiments with gp140 trimer showed that all of the antibodies tested had dissociation constants (K_d) ranging from 10^{-8} to 10^{-11} M; b12 performed at the lower end of the spectrum with a K_d of 1.2×10^{-8} M (Supplementary Fig. 7 and Supplementary Table 4). Thus, the IgG memory B cells obtained from the patients studied expressed high affinity antibodies specific for the CD4bs, the CD4i site and the VLs, and there was no immunodominant epitope.

In addition to anti-CD4bs, anti-CD4i and anti-VL antibodies, we found a group of antibodies that bound to gp120, gp120 core, gp120(D368R) and gp120(I420R) that we refer to as anti-gp120 core (18% of all gp140-binding antibodies; Supplementary Figs 8 and 9). These antibodies also bound to gp120(D368A/E370A) harbouring a double mutation that also interferes with binding to many of the known anti-CD4bs antibodies and CD4 (refs 17, 22). Furthermore, anti-gp120-core antibodies failed to bind to a stabilized gp120 core that is highly modified but retains CD4 and b12 antibody binding[23] (Supplementary Figs 8 and 9). However, none of the anti-gp120-core antibodies was sensitive to gp120 deglycosylation and therefore these antibodies are not predominantly directed to sugar moieties on gp120 (Supplementary Fig. 10).

To examine the properties of the anti-gp120-core antibodies further we performed inhibition ELISA experiments using biotin-labelled neutralizing antibodies to the CD4bs (b12 and 1–64), or CD4i (1–182), or the V3L (1–79) or a representative member of the gp120-core-specific group (2–491) (Figs 3d, 4 and Supplementary Tables 3, 5 and 6). Anti-gp120-core antibodies resembled b12 and CD4bs antibodies in that they inhibited the binding of the selected anti-gp120-core, anti-CD4bs, and anti-CD4i, but they did not inhibit binding of the anti-V3L antibody. Conversely, the 2–491 anti-gp120-core antibody was inhibited by the other anti-gp120-core and anti-CD4bs antibodies (Fig. 3d and Supplementary Tables 3, 5 and 6). However, only three out of thirteen of the anti-CD4i antibodies, and none of the seven anti-VL antibodies, inhibited binding of the anti-gp120 core (Fig. 3d and Supplementary Tables 3, 5 and 6). The affinity of anti-gp120-core antibodies to gp140 is similar to that of the anti-CD4bs antibodies (K_d values ranging from 2×10^{-8} to 4.8×10^{-10} M; Supplementary Table 4 and Supplementary Fig. 7). We conclude that anti-gp120-core antibodies recognize one or more immunogenic epitopes in the vicinity of the CD4bs and CD4i sites, but the precise targets for this group of antibodies on the HIV spike remains to be defined.

To determine the neutralizing activity of the memory antibodies we measured their ability to inhibit infection of TZM-bl cells by Env pseudovirus variants[24]. To determine whether there was intraclonal variation in neutralizing activity we also assayed somatic variants of some of the antibodies. Finally, purified serum IgG from the patients was assayed on the same viruses. The breadth of neutralizing activity and the relative sensitivity of different viral strains was similar for serum and purified IgG, indicating that most of the neutralizing activity was in the IgG fraction. Purified IgG neutralized most of the strains tested, but the activity was most pronounced for the more easily neutralized tier-1 HIV variants, whereas high concentrations of serum IgG were required for the more resistant strains (Fig. 4 and Supplementary Table 6).

Interestingly, 76% of all anti-gp120s and none of the anti-gp41s showed neutralizing activity at the concentrations tested (Fig. 4 and Supplementary Table 6). All anti-CD4bs and 88% of all anti-gp120-core antibodies showed some neutralizing activity (Fig. 4 and Supplementary Table 6). Of a total of 65 independent clonal families of neutralizing antibodies, 22 were anti-gp120(core), 18 were anti-CD4bs, 17 were anti-CD4i, and 8 were anti-VL including all three of the anti-V3L antibodies (Fig. 4 and Supplementary Table 6). As a group, the antibodies to the CD4bs and gp120 core showed the highest levels of activity with rare antibodies showing activity against the more resistant tier-2 viruses 6535.3 and SC422661.8 at high concentrations (Fig. 4 and Supplementary Table 6).

Although some degree of neutralizing activity was common among gp120-specific memory antibodies, we found no case in which a single monoclonal antibody accounts for all of the neutralizing activity in serum (Fig. 4 and Supplementary Table 6). Instead, individual antibodies showed variable levels of activity against different viruses. As a group these antibodies recognized a broad array of epitopes and neutralizing activity was heterogeneous for different viral isolates.

Memory B cells are long-lived cells that can differentiate into antibody-secreting plasma cells, but the relative contribution of any given memory B cell to the plasma cell compartment is unknown and

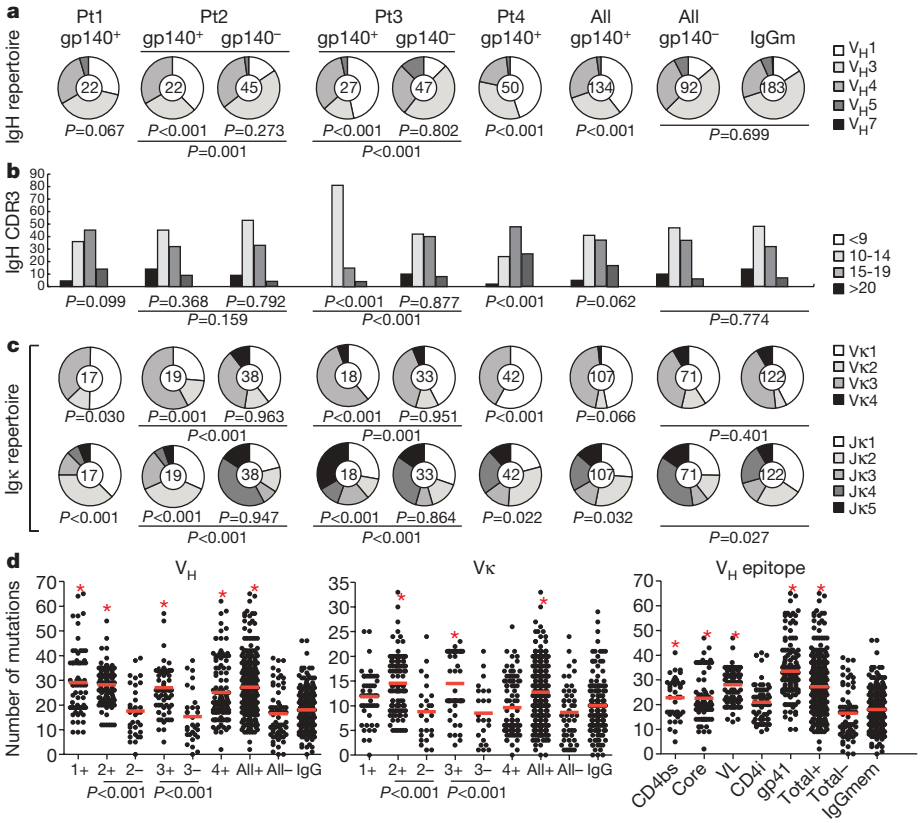


Figure 2 | Anti-gp140 antibody repertoire. Top line indicates patient (Pt) number and gp140 binding. IgGm are previously published controls[11]. Each clone is represented once irrespective of the clone size, or somatic variants. **a**, V_H repertoire analysis. **b**, IgH CDR3 length. **c**, Igκ repertoire comparing Vκ and Jκ. **d**, Graphs show the numbers of mutations per antibody for V_H and Vκ grouped by patient or V_H (right) grouped by epitope. Red asterisks indicate $P \leq 0.001$. *P* values were calculated by comparison to the pool of gp140 non-reactive antibodies except those below the lines, which refer to the paired samples.

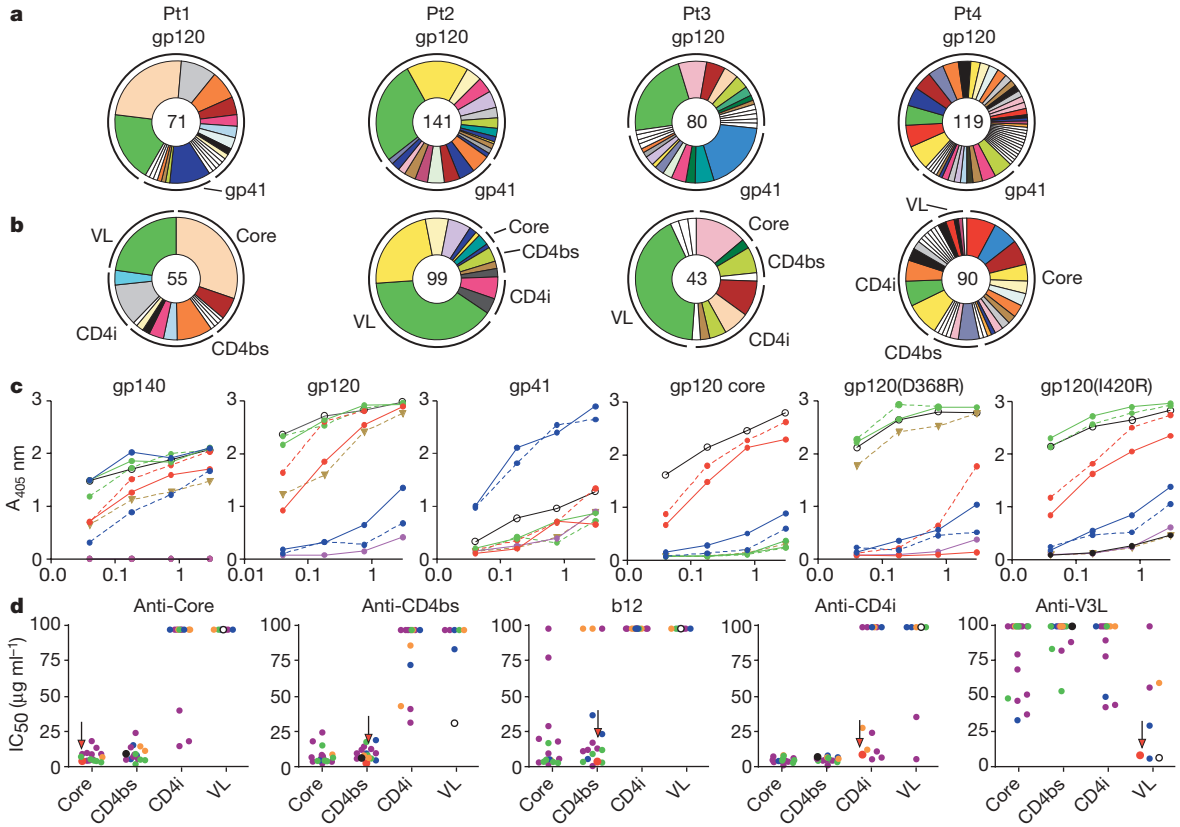


Figure 3 | Anti-gp140 mapping by ELISA. **a**, Pie charts show the distribution of anti-gp120 and anti-gp41 antibodies. **b**, Pie charts show the distribution of antibodies binding to CD4bs, CD4i, VL and gp120 core (Core). **c**, Representative ELISA results. The *x* axes show the antibody concentrations (in μg ml^{-1}). Green, 447-52D, anti-VL[21]; blue, 2F5 anti-gp41 (ref. 26); red, b12 anti-CD4bs[25]; purple, negative control[29]; black, anti-Core 4-221; dashed green lines, 2–59 anti-VL; dashed blue lines, 3–384 anti-gp41; dashed red lines, 2–1262 anti-CD4bs. **d**, Competition ELISA for binding to gp120. Green, patient 1; blue, patient 2; orange, patient 3; purple, patient 4. Each dot indicates the IC_{50} (Supplementary Table 5). Red arrow shows self-inhibitory activity. Filled circle, b12; open circle, 447-52D.

therefore a pool of cloned memory B-cell antibodies cannot be compared directly to serum. Nevertheless, we created pools of all antibodies for each individual patient and compared them to purified IgG for neutralization (Fig. 4 and Supplementary Table 6). The pools contained equal concentrations of each of the anti-gp140 clones irrespective of clone size, potential competition for epitope binding or neutralizing activity.

Purified IgGs neutralized nearly all of the tier-1 viruses, and the corresponding pools of the recombinant antibodies were also active against these viruses (Fig. 4 and Supplementary Table 6). In addition, some of the antibody pools neutralized viruses that were not neutralized by the IgG fraction (Fig. 4 and Supplementary Table 6).

In contrast, much higher concentrations of the patients' serum IgG were required for tier-2 neutralizing activity, ranging from 49 to 1,258 μg ml^{-1}. Consistent with the more stringent requirements for tier-2 neutralization, only the pooled monoclonal antibodies from patients 1 and 4 reconstituted this type of activity and only reached half-maximal inhibitory concentrations (IC_{50} values) at high concentrations (Fig. 4 and Supplementary Table 6). In conclusion, the memory antibody compartment contains a large mixture of anti-HIV neutralizing antibodies, combinations of which can reach the breadth of activity found in the serum but only at high concentrations.

Since the discovery of HIV, several monoclonal antibodies to the envelope protein have been produced by random cloning of heavy and light chains in phage display libraries or by selection of antibody-secreting hybridomas, but only a few highly active broadly neutralizing antibodies have been obtained[1–3]. Among these, b12 (ref. 25), 2F5 and 4E10 (refs 14, 26), and 2G12 (ref. 27) have received the greatest

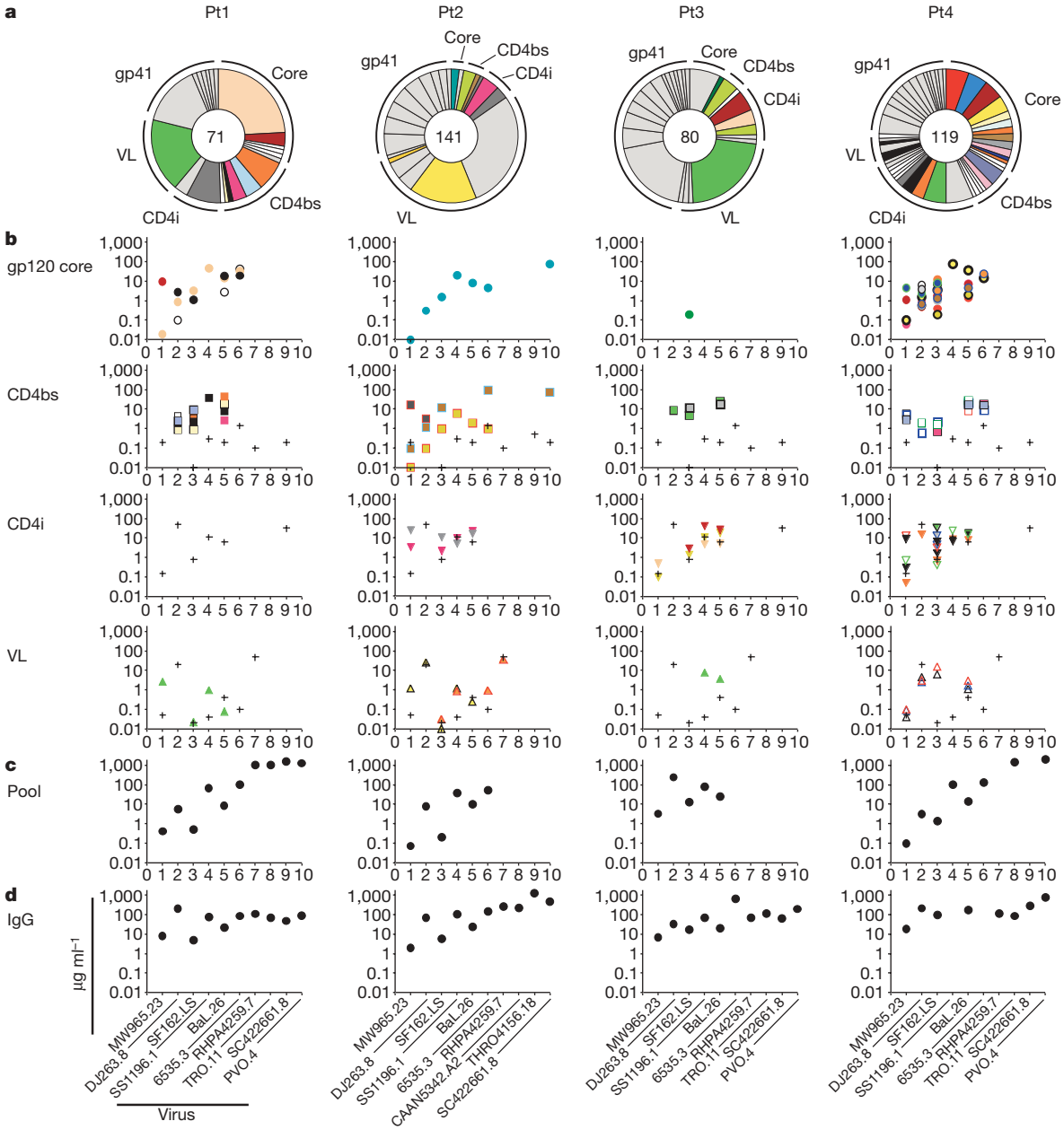


Figure 4 | Neutralizing activity. Patients (Pt1–Pt4) are indicated at the top. **a**, Pies show neutralizing antibodies in colour, non-neutralizers in grey. Epitopes are indicated. Slices are proportional to clone size. The number of antibodies is indicated in the centre. **b**, IC_{50} for individual antibodies to gp120 core, CD4bs, CD4i and VLs. The colours of the dots correspond to the pies above. Plus symbols indicate control antibodies b12 (CD4bs graphs), 17b (CD4i graphs) and 447-52D (VL graphs)[21]. **c**, Neutralizing activity of pooled anti-gp140s. **d**, Neutralizing activity of IgG. In **b**–**d**, the *y* axes show the antibody concentration (in μg ml^{-1}) required to achieve IC_{50}. The individual viruses on the *x* axes are indicated at the bottom (Supplementary Table 6).

attention because of their unique breadth and potency *in vitro* and *in vivo*. Ideally, a vaccine that induces such antibodies might be protective against HIV. However, to date, it has not been possible to re-isolate such antibodies from patients, or induce them by immunization in experimental animals[1–3]. Consistent with these findings, none of the 433 anti-gp140 antibodies we cloned from memory B cells from HIV-infected subjects with broad serum-neutralizing activity showed the type of broad activity exhibited by b12, 2F5, 4E10 or 2G12. Instead, the memory compartment contained many different neutralizing antibodies with more limited but diverse activity. Tier-2 neutralization was evident with mixtures of monoclonal antibodies but only at high concentrations. The molecular basis of this activity has not yet been determined, but it may result from the combination of positive additive effects of antibodies directed against different parts of gp140 and the negative effects of competition for binding to related epitopes by antibodies with high affinity but low neutralizing activity.

Our results do not rule out the possibility that broad neutralizing activity in serum can be the result of a single highly effective antibody, and the goal of eliciting such antibodies by vaccination remains important. However, the data suggest that a vaccine that phenocopies the natural anti-HIV immune response in patients with broadly neutralizing serological activity and elicits a combination of antibodies might also be an effective means of protection against a large number of HIV strains.

METHODS SUMMARY

Participants. HIV-1-infected patients are part of the Elite Controller Study of the Partners Aids Research Center (patients 2, 3 and 5) and clinical protocols at the Aaron Diamond Research Center (patient 1) and National Institute of Allergy and Infectious Diseases (patients 4 and 6) (Supplementary Table 1). The uninfected volunteers were recruited at the Rockefeller University. All work with human samples was performed in accordance with approved Institutional Review Board protocols.

Staining, single-cell sorting and antibody cloning. Staining and sorting of single gp140 binding memory B cells and cDNA cloning and expression was as previously described[8,9,28,29].

ELISA. Antigens were coated on 96-well plates as described[8]. For competition ELISAs, YU2-gp120-coated plates were incubated with pre-mixed biotinylated antibody and inhibiting antibody. Biotinylated antibody was detected using streptavidin-conjugated HRP (Serotec) and Horseradish Peroxidase Substrate Kit (Biorad).

Neutralization. Neutralization was measured as a reduction in luciferase reporter gene expression after single round infection in TZM-bl cells[24]. SIVmac251.WY5 and MuLV were used as negative controls to rule out nonspecific activity.

Received 25 January; accepted 27 February 2009.
Published online 15 March 2009.

1. Mascola, J. R. HIV/AIDS: allied responses. *Nature* **449,** 29–30 (2007).
2. Karlsson Hedestam, G. B. *et al.* The challenges of eliciting neutralizing antibodies to HIV-1 and to influenza virus. *Nature Rev. Microbiol.* **6,** 143–155 (2008).
3. Zolla-Pazner, S. Identifying epitopes of HIV-1 that induce protective antibodies. *Nature Rev. Immunol.* **4,** 199–210 (2004).
4. Shibata, R. *et al.* Neutralizing antibody directed against the HIV-1 envelope glycoprotein can completely block HIV-1/SIV chimeric virus infections of macaque monkeys. *Nature Med.* **5,** 204–210 (1999).
5. Mascola, J. R. *et al.* Protection of Macaques against pathogenic simian/human immunodeficiency virus 89.6PD by passive transfer of neutralizing antibodies. *J. Virol.* **73,** 4009–4018 (1999).
6. Trkola, A. *et al.* Delay of HIV-1 rebound after cessation of antiretroviral therapy through passive transfer of human neutralizing antibodies. *Nature Med.* **11,** 615–622 (2005).
7. Wei, X. *et al.* Antibody neutralization and escape by HIV-1. *Nature* **422,** 307–312 (2003).
8. Tiller, T. *et al.* Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. *J. Immunol. Methods* **329,** 112–124 (2008).
9. Scheid, J. F. *et al.* A method for identification of HIV gp140 binding memory B cells in human blood. *J. Immunol. Methods* doi:10.1016/j.jim.2008.11.012(in the press).
10. Mietzner, B. *et al.* Autoreactive IgG memory antibodies in patients with systemic lupus erythematosus arise from nonreactive and polyreactive precursors. *Proc. Natl Acad. Sci. USA* **105,** 9727–9732 (2008).
11. Tiller, T. *et al.* Autoreactivity in human IgG^+ memory B cells. *Immunity* **26,** 205–213 (2007).
12. Huang, C. C. *et al.* Structural basis of tyrosine sulfation and V_H-gene usage in antibodies that recognize the HIV type 1 coreceptor-binding site on gp120. *Proc. Natl Acad. Sci. USA* **101,** 2706–2711 (2004).
13. Muster, T. *et al.* A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1. *J. Virol.* **67,** 6642–6647 (1993).
14. Zwick, M. B. *et al.* Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. *J. Virol.* **75,** 10892–10905 (2001).
15. Xu, J. Y., Gorny, M. K., Palker, T., Karwowska, S. & Zolla-Pazner, S. Epitope mapping of two immunodominant domains of gp41, the transmembrane protein of human immunodeficiency virus type 1, using ten human monoclonal antibodies. *J. Virol.* **65,** 4832–4838 (1991).
16. Pantophlet, R. *et al.* Fine mapping of the interaction of neutralizing and nonneutralizing monoclonal antibodies with the CD4 binding site of human immunodeficiency virus type 1 gp120. *J. Virol.* **77,** 642–658 (2003).
17. Olshevsky, U. *et al.* Identification of individual human immunodeficiency virus type 1 gp120 amino acids important for CD4 receptor binding. *J. Virol.* **64,** 5701–5707 (1990).
18. Thali, M. *et al.* Characterization of a discontinuous human immunodeficiency virus type 1 gp120 epitope recognized by a broadly reactive neutralizing human monoclonal antibody. *J. Virol.* **65,** 6188–6193 (1991).
19. Thali, M. *et al.* Characterization of conserved human immunodeficiency virus type 1 gp120 neutralization epitopes exposed upon gp120–CD4 binding. *J. Virol.* **67,** 3978–3988 (1993).
20. Kwong, P. D. *et al.* Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393,** 648–659 (1998).
21. Gorny, M. K. *et al.* Neutralization of diverse human immunodeficiency virus type 1 variants by an anti-V3 human monoclonal antibody. *J. Virol.* **66,** 7538–7542 (1992).
22. Li, Y. *et al.* Broad HIV-1 neutralization mediated by CD4-binding site antibodies. *Nature Med.* **13,** 1032–1034 (2007).
23. Zhou, T. *et al.* Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature* **445,** 732–737 (2007).
24. Li, M. *et al.* Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *J. Virol.* **79,** 10108–10125 (2005).
25. Burton, D. R. *et al.* A large array of human monoclonal antibodies to type 1 human immunodeficiency virus from combinatorial libraries of asymptomatic seropositive individuals. *Proc. Natl Acad. Sci. USA* **88,** 10134–10137 (1991).
26. Buchacher, A. *et al.* Generation of human monoclonal antibodies against HIV-1 proteins; electrofusion and Epstein-Barr virus transformation for peripheral blood lymphocyte immortalization. *AIDS Res. Hum. Retroviruses* **10,** 359–369 (1994).
27. Trkola, A. *et al.* Human monoclonal antibody 2G12 defines a distinctive neutralization epitope on the gp120 glycoprotein of human immunodeficiency virus type 1. *J. Virol.* **70,** 1100–1108 (1996).
28. Yang, X., Farzan, M., Wyatt, R. & Sodroski, J. Characterization of stable, soluble trimers containing complete ectodomains of human immunodeficiency virus type 1 envelope glycoproteins. *J. Virol.* **74,** 5716–5725 (2000).
29. Wardemann, H. *et al.* Predominant autoantibody production by early human B cell precursors. *Science* **301,** 1374–1377 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Wycuff, E. Lybarger and B. Dey for supplying gp120 proteins for mapping studies, K. McKee for serum adsorption studies and N. Doria-Rose for her work with patient material from patients 4 and 6. This research was supported by the Rockefeller University, the International Aids Vaccine Initiative, the Bill and Melinda Gates Foundation, the Intramural Research Program of the Vaccine Research Center (R.T.W., J.R.M.), and the Division of Intramural Research (M.C.), National Institute of Allergy and Infectious Diseases, National Institutes of Health. J.F.S. was supported by the Deutscher Akademischer Austauschdienst, H.M. was supported by the Fondation Recherche Médicale. M.C.N. is a Howard Hughes Medical Institute investigator.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.C.N. (nussen@mail.rockefeller.edu).

LETTERS

Adaptation of HIV-1 to human leukocyte antigen class I

Yuka Kawashima[1], Katja Pfafferott[3,6], John Frater[4,5], Philippa Matthews[3], Rebecca Payne[3], Marylyn Addo[7], Hiroyuki Gatanaga[2,8], Mamoru Fujiwara[1], Atsuko Hachiya[1,8], Hirokazu Koizumi[1], Nozomi Kuse[1], Shinichi Oka[2,8], Anna Duda[4,5], Andrew Prendergast[3], Hayley Crawford[3], Alasdair Leslie[3], Zabrina Brumme[7], Chanson Brumme[7], Todd Allen[7], Christian Brander[7,9], Richard Kaslow[10], James Tang[10], Eric Hunter[11], Susan Allen[12], Joseph Mulenga[12], Songee Branch[13], Tim Roach[13], Mina John[6], Simon Mallal[6], Anthony Ogwu[14], Roger Shapiro[14], Julia G. Prado[3], Sarah Fidler[15], Jonathan Weber[15], Oliver G. Pybus[16], Paul Klenerman[4,5], Thumbi Ndung'u[17], Rodney Phillips[4,5], David Heckerman[19], P. Richard Harrigan[18], Bruce D. Walker[7,17,20], Masafumi Takiguchi[1] & Philip Goulder[3,6,17]

The rapid and extensive spread of the human immunodeficiency virus (HIV) epidemic provides a rare opportunity to witness host–pathogen co-evolution involving humans. A focal point is the interaction between genes encoding human leukocyte antigen (HLA) and those encoding HIV proteins. HLA molecules present fragments (epitopes) of HIV proteins on the surface of infected cells to enable immune recognition and killing by $CD8^+$ T cells; particular HLA molecules, such as HLA-B*57, HLA-B*27 and HLA-B*51, are more likely to mediate successful control of HIV infection[1]. Mutation within these epitopes can allow viral escape from $CD8^+$ T-cell recognition. Here we analysed viral sequences and HLA alleles from >2,800 subjects, drawn from 9 distinct study cohorts spanning 5 continents. Initial analysis of the HLA-B*51-restricted epitope, TAFTIPSI (reverse transcriptase residues 128–135), showed a strong correlation between the frequency of the escape mutation I135X and HLA-B*51 prevalence in the 9 study cohorts ($P = 0.0001$). Extending these analyses to incorporate other well-defined $CD8^+$ T-cell epitopes, including those restricted by HLA-B*57 and HLA-B*27, showed that the frequency of these epitope variants ($n = 14$) was consistently correlated with the prevalence of the restricting HLA allele in the different cohorts (together, $P < 0.0001$), demonstrating strong evidence of HIV adaptation to HLA at a population level. This process of viral adaptation may dismantle the well-established HLA associations with control of HIV infection that are linked to the availability of key epitopes, and highlights the challenge for a vaccine to keep pace with the changing immunological landscape presented by HIV.

The extent to which HIV is evolving at the population level in response to immune selection pressure is under debate[2–6]. Resolving the impact of HLA class I alleles on viral evolution is problematic because it can be obscured by other influences, such as founder effect[6] (polymorphisms present within the early strains establishing the epidemic in a group). In addition, most HLA alleles do not drive significant selection pressure on HIV, a proportion of escape mutations revert to wild type after transmission, and different HLA alleles may drive the identical escape mutation[7].

To test the hypothesis that the frequency of escape mutations in a given population is correlated with the prevalence of the relevant HLA allele in that population, we studied nine distinct cohorts from North America, the Caribbean, Europe, sub-Saharan Africa, Australia and Japan, in which we performed HLA typing, and defined the viral mutations arising within $CD8^+$ T-cell epitopes. We focused initially on a well-characterized mutation, I135X, within the HLA-B*51-restricted epitope, TAFTIPSI (RT 128–135)[8], because it arises in acute infection, non-HLA-B*51 alleles do not also select this mutation[7,9], and it does not revert to Ile 135 after transmission to HLA-B*51-negative subjects[9]. Thus, if highly prevalent HLA alleles drive a high frequency of escape mutations in the population, this would be most obvious in relation to HLA-B*51 and the escape mutant I135X. We then considered an additional 13 well-defined escape mutations, including those known to reduce viral fitness and therefore liable to revert after transmission.

I135X was selected in 205 of 213 (96%) HLA-B*51-positive individuals analysed (Figs 1 and 2, and Supplementary Fig. 1). The I135X variants do not significantly affect viral replicative capacity *in vitro*, other than the rare I135V mutation. This was the only variant observed to revert to wild-type *in vivo* during a 3-year follow-up of 38 HLA-B*51-negative subjects identified during acute HIV infection who carried I135X mutant viruses at transmission (Fig. 1e). The I135X mutants substantially affect HLA binding, and therefore also recognition by $CD8^+$ T cells (Fig. 1f–h). Thus, HIV transmission from HLA-B*51-positive subjects would probably involve transmission of I135X, which would persist in the new host. Newly infected HLA-B*51-positive subjects receiving an I135X mutant would be unable to generate an HLA-B*51-TAFTIPSI-specific response.

To test the hypothesis that the population frequency of I135X is correlated with HLA-B*51 prevalence, HIV sequence and HLA data were collated from the nine study cohorts. One cohort comprised subjects with acute/early HIV infection; the remaining cohorts comprised chronically infected subjects. In all cohorts the odds ratio strongly favoured I135X in the HLA-B*51-positive subjects, even in the acute cohort where I135X was selected sufficiently early to be already over-represented in HLA-B*51-positive subjects (odds ratio 1.65, $P = 0.07$, Fig. 2a). In Japan, where HLA-B*51 is highly

[1]Divisions of Viral Immunology and [2]Infectious Disease, Center for AIDS Research, Kumamoto University, 2-2-1 Honjo, Kumamoto 860-0811, Japan. [3]Department of Paediatrics, [4]Nuffield Department of Clinical Medicine and [5]The James Martin 21st Century School, Peter Medawar Building for Pathogen Research, South Parks Road, Oxford OX1 3SY, UK. [6]Centre for Clinical Immunology and Biomedical Statistics, Royal Perth Hospital and Murdoch University, Western Australia 6000, Australia. [7]Partners AIDS Research Center, Massachusetts General Hospital, 13th Street, Building 149, Charlestown, Boston, Massachusetts 02129, USA. [8]AIDS Clinical Center, International Medical Center of Japan, 1-21-1 Toyama, Shinjuku-ku, Tokyo 162-8655, Japan. [9]Fundació IrsiCaixa-HIVACAT, Hospital Germans Trias i Pujol, Badalona and Institucio Catalana de Recerca i Estudis Avancats (ICREA), Barcelona 08916, Spain. [10]University of Alabama at Birmingham, Birmingham, Alabama 35294, USA. [11]Emory University Vaccine Center and Yerkes National Primate Research Center, Atlanta, Georgia 30329, USA. [12]Zambia Emory HIV Research Project, and the Zambia Blood Transfusion Service, Lusaka, Zambia. [13]Ladymeade Reference Unit, University of West Indies, Bridgetown BB11156, Barbados. [14]Botswana-Harvard School of Public Health AIDS Initiative Partnership, Gaborone, Botswana. [15]Division of Medicine, Wright Fleming Institute, Imperial College, St Mary's Hospital, Norfolk Place, Paddington, London W2 1PG, UK. [16]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3SY, UK. [17]HIV Pathogenesis Programme, The Doris Duke Medical Research Institute, University of KwaZulu-Natal, Durban 4013, South Africa. [18]Miscrosoft Research, One Microsoft Way, Redmond, Washington 9805, USA. [19]BC Centre for Excellence in HIV/AIDS, Vancouver, British Columbia V6Z 1Y6, Canada. [20]Howard Hughes Medical Institute, Chevy Chase, Maryland 20185, USA.

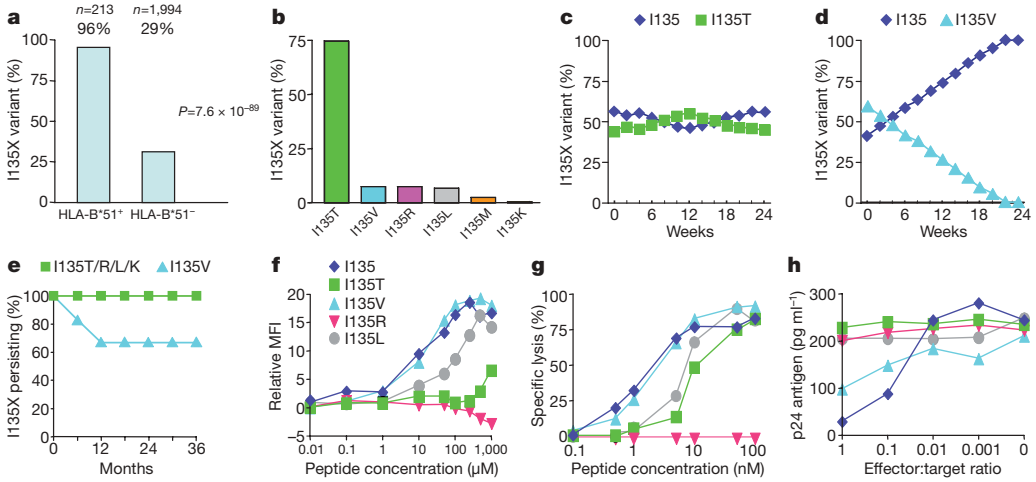


Figure 1 | Selection and fitness cost of I135X escape variants and recognition by the HLA-B*51–TAFTIPSI (RT 128–135)-specific CD8+ T cells. **a**, Association between I135X and HLA-B*51 in all study cohorts. **b**, Ile 135 variation in HLA-B*51-positive subjects. **c**, **d**, *In vitro* competition assays between NL4-3 wild-type virus and I135X viral variants (I135T (**c**) and I135V (**d**)). I135R and I135L showed no fitness cost (not shown). **e**, Persistence of I135X mutants in 38 HLA-B*51-negative subjects followed from acute infection. **f**, TAFTIPSI variant binding to HLA-B*51 (see Methods). MFI, mean fluorescence intensity. **g**, **h**, Recognition of peptide-pulsed HLA-B*51-matched targets and viral variants by representative TAFTIPSI-specific CD8+ T-cell clones.

prevalent[10] (21.9% of the study cohort), the frequency of I135X was >50%, and overall across all cohorts the I135X frequency was strongly correlated with HLA-B*51 prevalence ($P = 0.0001$, Fig. 2b). To control for the possibility that disproportionately more virus sequences from HLA-B*51-positive subjects were analysed, the same analysis comparing I135X frequency in HLA-B*51-negative subjects only was undertaken, with similar findings (Fig. 2c, $P = 0.0006$). These data suggest that HIV may be adapting to HLA-B*51 with respect to the HLA-B*51–TAFTIPSI response in localities where HLA-B*51 is at high prevalence.

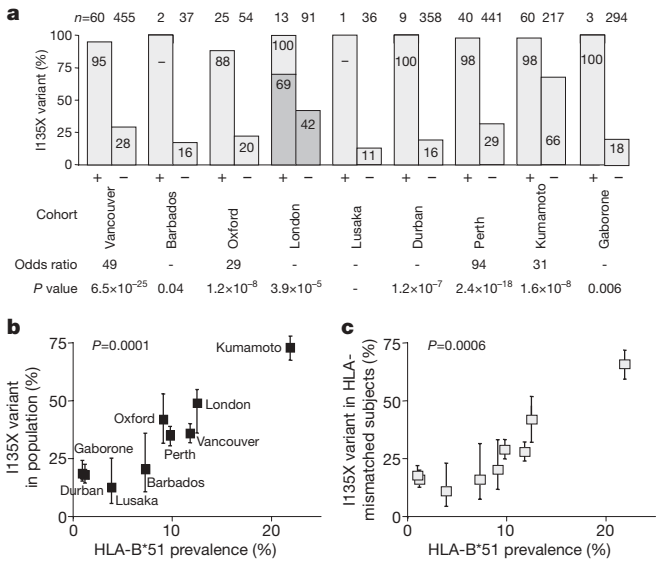


Figure 2 | Correlation between frequency of HLA-B*51-associated escape mutations and HLA-B*51 prevalence in study cohorts. **a**, Frequency of I135X mutations within TAFTIPSI (RT 128–135) in HLA-B*51-positive (+) and -negative (−) subjects within nine study cohorts. In the acute cohort (London) 69% of HLA-B*51-positive subjects expressed I135X mutant at enrolment, 100% within 2 years of baseline (Supplementary Fig. 1). **b**, Correlation between frequency of I135X mutation and HLA-B*51 prevalence in the nine study populations. Logistic regression $P = 0.0001$ (Supplementary Table 1). **c**, Correlation between I135X frequency in HLA-B*51-negative subjects and HLA-B*51 prevalence in nine study populations. Error bars represent 95% confidence limits, obtained using a binomial error distribution.

Additional evidence that I135X is accumulating in Japan comes from the observation that only 3 of 14 (21%) HLA-B*51-negative Japanese haemophiliacs infected in 1983 carried I135X, compared with 30 of 43 (70%) HLA-B*51-negative subjects infected between 1997 and 2008 ($P = 0.002$). Furthermore, HLA-B*51 does not protect against disease progression in Japanese subjects infected between 1997 and 2008, whereas HLA-B*51-positive haemophiliacs infected in 1983 had lower viraemia levels and higher CD4 counts than HLA-B*51-negative haemophiliacs (Supplementary Fig. 2). These data are consistent with fewer HLA-B*51-positive subjects targeting TAFTIPSI during 1997–2008, owing to a population-level increase in the HLA-B*51 I135X escape mutation over this 14–25-year period.

To investigate HIV adaptation to other HLA alleles, we initially examined other escape mutations shown previously to persist stably after transmission[5,7]. We selected the three non-reverting Gag polymorphisms that, from analysis of 673 study subjects in Durban, South Africa[7], were most strongly associated with the relevant restricting allele ($P < 10^{-6}$ after phylogenetic correction), namely, S357X, D260X and D312X within epitopes restricted, respectively, by HLA-B*07 (GPSHKARVL, Gag 355–363), HLA-B*35 (PPIPVGDIY, Gag 254–262) and HLA-B*44 (AEQATQDVKNW, Gag, 306–316). In addition, we analysed a non-reverting I31V variant (LPPIVAKEI, Int 28–36) previously hypothesized to increase in relation to population HLA-B*51 prevalence[5]. These additional polymorphisms show a similar relationship to that between I135X and HLA-B*51, overall showing a strongly significant correlation between variant frequency and prevalence of the restricting HLA allele (Figs 3 and 4a, and Supplementary Fig. 3).

The spectrum of HLA-associated polymorphisms also includes mutations reducing viral fitness[1]. These either revert to wild type after transmission, or persist in the presence of compensatory mutations. We extended these analyses to include epitopes restricted by HLA-B*27 and HLA-B*57, alleles strongly associated with successful immune control of HIV[11,12]. The mutations analysed themselves are associated with precipitating loss of immune control[13–16] and all inflict a documented viral fitness cost, either demonstrated by *in vitro* fitness studies and/or *in vivo* reversion[7,14,17–21] (data not shown for V168I).

Again, a strong correlation between escape mutant frequency and prevalence of the restricting HLA allele was observed (Figs 3c–f and 4b, and Supplementary Fig. 3; overall, for these nine variants affecting viral fitness, $r = 0.69$, $P < 0.0001$). Unexpectedly, this correlation

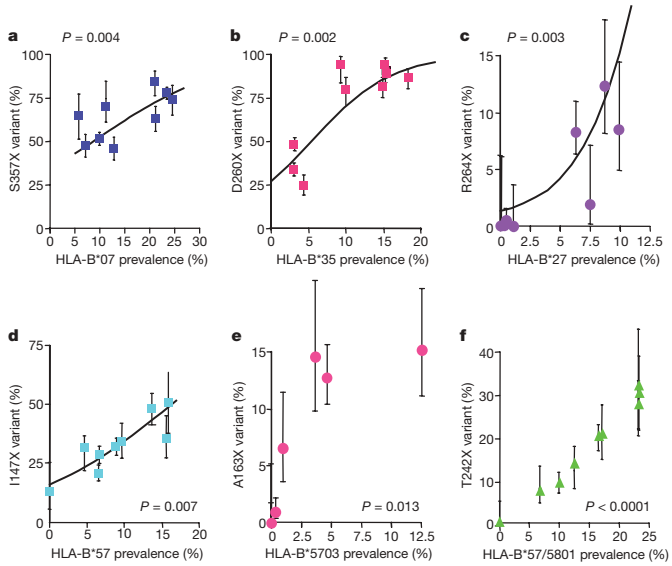


Figure 3 | Correlation between frequency of HIV sequence variant and HLA prevalence for six additional well-characterized epitopes. *P* values calculated after logistic regression analysis as shown (calculations after linear regression analysis are shown in Supplementary Table 1). **a**, Frequency of the S357X mutation within the HLA-B*07-restricted epitope GPSHKARVL (Gag 355–363). **b**, Frequency of the D260X mutation within the HLA-B*35-restricted epitope PPIPVGDIY (Gag 254–262). **c**, Frequency of the R264X mutation within the HLA-B*27-restricted epitope KRWIILGLNK (Gag 263–272). **d**, Frequency of the I147X mutation within the HLA-B*57-restricted epitope ISPRTLNAW (Gag 147–155). **e**, Frequency of the A163X mutation associated with the HLA-B*5703-restricted epitope KAFSPEVIPMF (Gag 162–172). **f**, Frequency of the T242X mutation within the B*57/5801-restricted epitope TSTLQEQIAW (Gag 240–249). Error bars represent 95% confidence limits, obtained using a binomial error distribution.

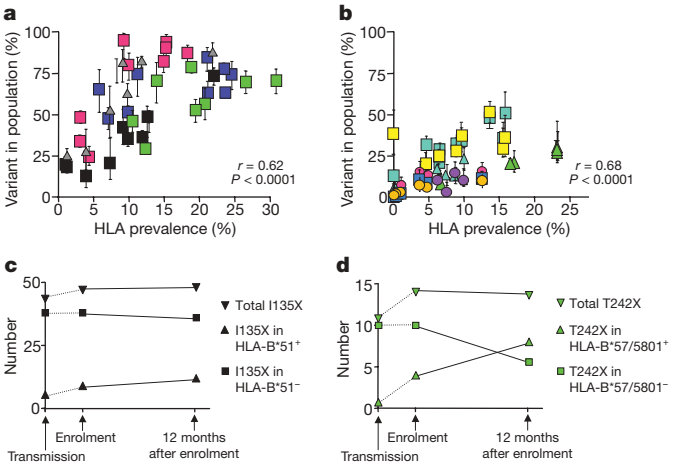


Figure 4 | Correlation between HIV variant frequency and HLA prevalence for all epitopes studied. **a**, Correlation between HLA prevalence and the five stable, non-reverting variants (symbols in Figs 2 and 3, and Supplementary Fig. 3; grey triangles, I31V; green squares, D312X). **b**, Eight variants demonstrated to reduce viral fitness (see text, Fig. 3 and Supplementary Fig. 3; turquoise triangles, L268X; yellow squares. A146X; sky-blue squares, V168I; yellow circles, I247X). **c**, **d**, Data from acute London cohort. **c**, Number of HLA-B*51-positive and HLA-B*51-negative subjects carrying the non-reverting I135X variant. The percentage of I135X in HLA-B*51-negative subjects at enrolment (42%) assumed the percentage of I135X in all subjects at transmission (I135X frequency in HLA-B*51-positive subjects at enrolment was 69%, $P = 0.07$). **d**, The reverting HLA-B*57/5801-restricted T242X mutation. T242X frequency in HLA-B*57/5801-negative subjects at enrolment was 7%, versus 33% in HLA-B*57/5801-positive subjects ($P = 0.01$). Error bars represent 95% confidence limits, obtained using a binomial error distribution.

remained significant even when comparing HLA prevalence with variant frequency in the HLA-mismatched population ($r = 0.40$, $P = 0.0004$). As anticipated, non-reverting variants such as I135X accumulate at the population level, but even rapidly reverting[18,20] mutations such as T242N can accumulate, if the selection rate exceeds the reversion rate (Fig. 4c, d).

Although frequency of the analysed HIV polymorphisms and HLA prevalence were strongly correlated overall, some anomalies were observed. For example, despite a 0% prevalence of HLA-B*57 in Japan[10], 38% of the Japanese cohort had the HLA-B*57-associated A146X variant. One potential explanation might be A146X selection by non-HLA-B*57 Japanese alleles. Analysing Gag sequences from Japanese study subjects, we observed a strong association between A146P and HLA-B*4801 ($P = 0.00035$), and then that A146P is indeed selected in HLA-B*4801-positive subjects (Supplementary Fig. 4a, b). We defined a novel HLA-B*4801-restricted epitope (Gag 138–147), showing also that A146P is an escape mutant (Supplementary Fig. 4c–f). These data illustrate that more than one HLA allele can drive the selection of a particular escape mutant (Supplementary Fig. 5). Also, in populations where HIV-specific $CD8^+$ T-cell responses are incompletely characterized, the influences of locally prevalent HLA alleles on HIV sequence variation are unknown.

These data show a strong correlation between HLA-associated HIV sequence variation and HLA prevalence in the population ($r = 0.69$, $P < 0.0001$, Supplementary Fig. 6), suggesting that the frequency of the studied variants is substantially driven by the HLA-restricted $CD8^+$ T-cell responses. Non-reverting variants[5,7], as well as those previously shown to arise at a fitness cost[7,14,16–21], were studied. The latter constitute approximately 55–65% of HLA-associated polymorphisms[7,20]. This current analysis included epitopes whose role in HIV immune control is unknown, as well as those believed to contribute significantly to containment of HIV[4,7,13–19]. Analysis of well-characterized epitopes only also served to limit potential confounding influences of epitope clustering (selection of the same variant by different HLA alleles) and of founder effect. Either would be capable of obscuring a true HLA effect on population variant frequency.

The HLA-B*57-associated A146X mutation illustrates the complexity that may result from epitope clustering. A146X is selected by at least six distinct HLA alleles (Supplementary Fig. 5). A true correlation existing between mutation frequency and individual HLA allele prevalence might thus be obscured by selection of the same mutation by other alleles.

Founder effect also has an undoubted influence on population frequencies of particular polymorphisms[6]. Phylogenetic correction of sequence data excludes founder effect as a confounder[6,7,9], and the highly significant associations between the presence of particular HLA alleles and all 14 HIV polymorphisms studied, persisting after phylogenetic correction (Supplementary Table 3), provide compelling evidence that the effects observed here are substantially HLA-driven. The large numbers of study subjects in these current studies reduce the likelihood of genuine HLA associations with HIV amino acid polymorphisms being obscured by founder effects. The relative impact of HLA and founder effect on variant frequency is harder to quantify, and is likely to differ substantially between particular populations.

The consequence of HIV adapting to certain $CD8^+$ T-cell responses is unknown. For non-reverting polymorphisms such as HLA-B*35-associated D260E, the variant approaches fixation, because even at population frequencies of 90%, D260E is still significantly selected in HLA-B*35-positive subjects (Supplementary Fig. 7b). Important questions relevant to vaccine design include the extent and rate of sequence change in populations. Relevant factors include the selection rate in subjects expressing the HLA allele, the reversion rate in HLA-mismatched subjects, the population HIV

transmission rate, and HLA allele prevalence. Models would need to include factors such as the selection of compensatory mutations to slow reversion rates, and antiretroviral therapy access that would slow transmission rates.

HLA adaptation to certain $CD8^+$ T-cell responses may also alter currently established HLA associations with slow disease progression. Data here suggest that, whereas 25 years ago HLA-B*51 was protective in Japan[11,12], this is no longer the case (Supplementary Fig. 2). The apparent increase in I135X frequency in Japan over this time supports the notion that HLA-B*51 protection against HIV disease progression hinges on availability of the HLA-B*51-restricted TAFTIPSI response. However, whether this is the case remains unknown.

For HLA-B*27 and HLA-B*57, there is more clear-cut evidence that their association with HIV control depends on the Gag-specific epitopes presented and analysed here[4,7,13–15,18,19]. For each of the HLA-B*27- and HLA-B*57-associated Gag mutations studied, an *in vitro* fitness cost or *in vivo* reversion has been observed. A strong correlation between variant frequency and HLA prevalence even for rapidly reverting variants can be explained, either by mutant acquisition exceeding reversion rate (Fig. 4D), or by selection of compensatory mutations slowing or halting reversion altogether. The clearest example of the latter is the HLA-B*27-associated R264K mutation, 'corrected' by S173A[19]. Compensatory mutations are also well described for the HLA-B*57-associated Gag mutations[14,18]. These data suggest that the escape mutations in these HLA-B*27- and HLA-B*57-restricted epitopes are accumulating over time. Several studies have now demonstrated that transmission of viruses encoding escape mutants in the critical Gag epitopes to individuals expressing the relevant MHC class results in failure to control viraemia[2,21,22]. The accumulation at the population level of these escape mutations in HLA-B*27 and HLA-B*57 Gag epitopes is therefore likely to reduce the facility of these alleles to slow HIV disease progression.

The longer-term consequences of this process for immune control of HIV are unknown. Loss of currently immunodominant epitopes would promote subdominant $CD8^+$ T-cell responses, which can be more effective[23,24]. Also, the adapted virus provides new epitopes that can be presented, potentially with beneficial effects. In hepatitis C virus, for example, HLA-A*0301 holds a particular advantage, but only against the specific strain of virus responsible for the Irish outbreak[25]. In HIV, HLA-B*1801 is associated with high viraemia in C clade but not in B clade infection[10,11,26]; the opposite applies to HLA-B*5301.

Thus, the data presented here, showing evidence that the virus is adapting to $CD8^+$ T-cell responses, some of which may mediate the well-established associations (HLA-B*57, HLA-B*27 and HLA-B*51) with immune control of HIV, highlight the dynamic nature of the challenge for an HIV vaccine. Important questions to be addressed include the speed and extent of sequence change, particularly in Gag, the most effective target for $CD8^+$ T-cell responses[1,7,13,21]. The induction of broad Gag-specific $CD8^+$ T-cell responses may be a successful vaccine strategy, but such a vaccine will be most effective if tailored to the viral sequences prevailing, and thus may need to be modified periodically to keep pace with the evolving virus. Moreover, the strong associations between certain HLA class molecules, such as HLA-B*57, HLA-B*27 and HLA-B*51, and slow disease progression may decline as the epidemic continues, particularly where these HLA alleles are highly prevalent, and where HIV transmission rates are high.

METHODS SUMMARY

Overall 2,875 subjects were studied, from 9 previously established study cohorts. These cohorts comprised subjects from North America, the Caribbean, Europe, sub-Saharan Africa, Australasia and Asia. All subjects were antiretroviral-therapy-naive. Apart from the London acute cohort ($n = 142$), all cohorts comprised chronically infected subjects. The 14 variants studied are well-defined escape mutations within well-characterized $CD8^+$ T-cell epitopes, and included those persisting after transmission and likely to have little effect on viral fitness ($n = 5$), as well as those shown previously to reduce viral fitness ($n = 9$). Autologous HIV-1 sequences, and HLA class I types, were determined for all study subjects. The replicative capacity of I135X variants selected within the HLA-B*51-restricted epitope TAFTIPSI (RT 128–135) was assessed via *in vitro* competition assays and also via longitudinal follow-up of HLA-B*51-negative subjects infected acutely with I135X variants. Polymorphism frequency in the study cohorts was compared with prevalence of the relevant HLA molecule in the study cohort using a logistic regression model taking into account the different numbers of study subjects in each cohort. Demonstration of an HLA allele driving escape at Gag 146 in the Japanese cohort was undertaken first by identification of an association between HLA-B*4801 and A146P, subsequent definition of an HLA-B*4801-restricted $CD8^+$ T-cell response to a novel epitope Gag 138–147 (LI10), and finally demonstration that A146P reduced viral recognition by LI10-specific $CD8^+$ T cells.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 October; accepted 22 December 2008.
Published online 25 February 2009.

1. Goulder, P. J. R. & Watkins, D. I. Impact of MHC class I diversity on immune control of immunodeficiency virus replication. *Nature Rev. Immunol.* **8,** 619–630 (2008).
2. Goulder, P. J. R. *et al.* Evolution and transmission of stable CTL escape mutations in HIV infection. *Nature* **412,** 334–338 (2001).
3. Moore, C. B. *et al.* Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* **296,** 1439–1443 (2002).
4. Draenert, R. *et al.* Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *J. Exp. Med.* **199,** 905–915 (2004).
5. Leslie, A. J. *et al.* Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA. *J. Exp. Med.* **201,** 891–902 (2005).
6. Bhattacharya, T. *et al.* Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* **315,** 1583–1586 (2007).
7. Matthews, P. *et al.* Central role of reverting mutations in HLA associations with viral setpoint. *J. Virol.* **82,** 8548–8559 (2008).
8. Tomiyama, H. *et al.* Identification of multiple HIV-1 CTL epitopes presented by HLA-B*5101 molecules. *Hum. Immunol.* **60,** 177–186 (1999).
9. Brumme, Z. *et al.* Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection. *AIDS* **22,** 1277–1286 (2008).
10. Itoh, Y. *et al.* High throughput DNA tying of HLA-A, -B, -C, and -DRB1 loci by a PCR-SSOP-Luminex method in the Japanese population. *Immunogenetics* **57,** 717–729 (2005).
11. Kaslow, R. A. *et al.* Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nature Med.* **2,** 405–411 (1996).
12. O'Brien, S. J., Gao, X. & Carrington, M. HLA and AIDS: a cautionary tale. *Trends Mol. Med.* **7,** 379–381 (2001).
13. Kiepiela, P. *et al.* $CD8^+$ T-cell responses to different HIV proteins have discordant associations with viral load. *Nature Med.* **13,** 46–53 (2007).
14. Leslie, A. J. *et al.* HIV evolution: CTL escape mutation and reversion after transmission. *Nature Med.* **10,** 282–289 (2004).
15. Goulder, P. J. R. *et al.* Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nature Med.* **3,** 212–217 (1997).
16. Feeney, M. E. *et al.* Immune escape precedes breakthrough HIV-1 viremia and broadening of the CTL response in a HLA-B27-positive long-term nonprogressing child. *J. Virol.* **78,** 8927–8930 (2004).
17. Martinez-Picado, J. *et al.* Fitness cost of escape mutations in p24 Gag in association with control of human immunodeficiency virus type 1. *J. Virol.* **80,** 3617–3623 (2006).
18. Crawford, H. *et al.* Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted Gag epitope in chronic human immunodeficiency virus type 1 infection. *J. Virol.* **81,** 8346–8351 (2007).
19. Schneidewind, A. *et al.* Escape from a dominant Gag-specific CTL response in HLA-$B27^+$ subjects is associated with a dramatic reduction in HIV-1 replication. *J. Virol.* **81,** 12382–12393 (2007).
20. Brumme, Z. *et al.* Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 (HIV-1) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection. *J. Virol.* **82,** 9216–9227 (2008).
21. Goepfert, P. *et al.* Transmission of Gag immune escape mutations is associated with reduced viral load in linked recipients. *J. Exp. Med.* **205,** 1009–1017 (2008).
22. Seki, S. *et al.* Transmission of SIV carrying multiple cytotoxic T lymphocyte escape mutations with diminished replicative capacity can result in AIDS progression in Rhesus macaques. *J. Virol.* **82,** 5093–5098 (2008).

23. Gallimore, A., Dumrese, T., Hengartner, H., Zinkernagel, R. M. & Rammensee, H. G. Protective immunity does not correlate with the hierarchy of virus-specific cytotoxic T cell responses to naturally processed peptides. *J. Exp. Med.* **187,** 1647–1657 (1998).
24. Holtappels, R. *et al.* Subdominant CD8 T-cell epitopes account for protection against cytomegalovirus independent of immunodomination. *J. Virol.* **82,** 5781–5796 (2008).
25. McKiernan, S. M. *et al.* Distinct MHC class I and II alleles are associated with hepatitis C viral clearance, originating from a single source. *Hepatology* **40,** 108–114 (2004).
26. Kiepiela, P. *et al.* Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432,** 769–775 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work is funded by grants from the National Institutes of Health (RO1AI46995 (P.G.), 1 R01 AI067073 (B.D.W.), R01AI64060 (E.H.)), the Wellcome Trust (P.G., P.K.), the UK Medical Research Council (J.F., A.P. and P.M.), and the Mark and Lisa Schwartz Foundation, the Ministry of Health, Labour and Welfare (Health and Labour HIV/AIDS Research Grants 012), the NIHR Biomedical Research Centre Programme and the Ministry of Education, Science, Sports and Culture (number 18390141), Japan (M.T.). P.G. is an Elizabeth Glaser Pediatric AIDS Foundation Scientist; J.G.P. is a Marie Curie Fellow (contract number IEF-041811). The authors are also grateful to A. McLean and H. Fryer for discussions of the manuscript.

Author Contributions Y.K., K.P., J.F. and P. M. undertook much of the experimental work and data analysis, and contributed equally. M.T. and P.G. undertook much of the project conception, planning, supervision, analysis and writing of the manuscript, and contributed equally.

Author Information Accession numbers for newly determined viral sequences are included in Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.G. (philip.goulder@paediatrics.ox.ac.uk).

METHODS

Study subjects. The study cohorts have been described more fully elsewhere[3,7,9,13,14,18,20,21,27]. All comprise chronically infected and highly active antiretroviral therapy (HAART)-naive study subjects, with the exception of the London acute cohort ($n = 142$), who were enrolled immediately after seroconversion between 1999 and 2004, and 54 subjects enrolled during acute infection in Japan between 1997 and 2008. Viral sequences in all 2,679 chronically infected study subjects (all of whom were HAART-naive) were determined from time points after 2000, with the exception of 9 study subjects in the Japanese chronic cohort (1998–99) and all of the British Columbia cohort (1996–99). Sequencing data were obtained from 566 study subjects in the British Columbia cohort, 53 study subjects in the Barbados cohort, 106 in the Oxford cohort, 673 in the Durban cohort, 226 in the Lusaka cohort (chronically infected subjects enrolled between 2005–08), 481 study subjects in the Perth cohort, 277 chronically infected subjects in the Kumamoto cohort, 297 in the Gaborone cohort, and 142 subjects in the acute London cohort. An additional cohort in Japan comprised 117 haemophiliacs who were infected before 1985, the majority of which were believed to have been infected in 1983, and who were enrolled and followed up in out-patient clinics since 1997. These haemophiliacs are all now on HAART except for 4 HAART-naive subjects.

HLA-associated HIV amino acid polymorphisms studied. Variants studied that were shown to reduce viral fitness comprised polymorphisms within the HLA-B*27-restricted Gag epitope KRWIILGLNK (Gag 263–272; R264X and L268X) and mutations in three HLA-B*57-restricted Gag epitopes: ISPRTLNAW (ISW9, Gag 147–155), KAFSPEVIPMF (KF11, Gag 162–172) and TSTLQEQIAW (TW10, Gag 240–249). T242X is strongly selected by HLA-B*5801 in addition to HLA-B*57 subtypes[7,14,17]. The HLA-B*57-associated polymorphisms at residues Gag 146, 147 and 248 are selected by all HLA-B*57 subtypes, whereas Gag 163, 165, 166 and 247 are only selected by the HLA-B*5703 subtype (refs 7, 18 and H.C., unpublished data).

Statistics. Polymorphism frequency in the study cohorts was compared with prevalence of the relevant HLA molecule in the study cohort using a logistic regression model. To take account of the different numbers of study subjects in each cohort, appropriate confidence limits for the mutation frequencies were calculated, using the Adjusted–Wald method for binomial variables[28]. Logistic regression was calculated by GLMStat (http://www.glmstat.com) using a binomial error distribution and a logit link function. In addition, the Spearman's rank correlation coefficient was calculated in the context of a linear regression model (data shown in Supplementary Tables 1 and 2).

HLA class I typing. Because HLA typing was not undertaken consistently to four-digit resolution in all cohorts, two-digit HLA types only were used for these analyses, with the exception of the HLA-B*5703-associated polymorphisms (the Barbados and Oxford cohorts being excluded from these latter analyses as HLA-B*57 subtyping data were not available). Genomic DNA samples were initially typed to an oligo-allelic (two-digit) level using Dynal RELITM reverse SSO kits for the HLA-A, HLA-B and HLA-C loci (Dynal Biotech). Refining the genotype to the allele level was performed using Dynal Biotech sequence-specific priming (SSP) kits in conjunction with the previous SSO type. HLA phenotypic frequencies were determined from the HIV-infected study cohorts themselves.

Sequencing of viral RNA and proviral DNA. Viral sequencing of *gag* and *pol* from plasma RNA and proviral DNA was undertaken, using primers as previously described[7,9]. PCR products were sequenced directly or they were cloned by using a TOPO TA cloning kit (Invitrogen) and then sequenced. Sequencing was done with a Big Dye terminator v1.1. cycle sequencing kit (Applied Biosystems) and analysed by an ABI PRISM 310 genetic Analyser.

Competitive HIV-1 replication assay. Freshly prepared H9 cells (3×10^5) were exposed to the mixtures of paired virus preparations (300 blue cell-forming units) each of NL-432 versus mutant virus (I135T, I135V, I135R and I135L)), to be examined for their replication ability for 2 h, washed twice with PBS, and cultured as described previously[29]. On day 1, one-third of infected H9 cells were harvested and washed twice with PBS, and the proviral HIV-1 reverse transcriptase gene was sequenced (0 week). Every 7 days, the supernatant of the virus culture was transmitted to new uninfected H9 cells. The cells harvested at the end of every other passage (that is, at 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 and 24 weeks) were subjected to direct DNA sequencing of the HIV-1 reverse transcriptase gene, and the viral population change was determined by the relative peak height on the sequencing electrogram. The persistence of the original amino acid substitution was confirmed for all infectious clones used in this assay.

HLA-B*5101 stabilization assay. Binding of HIV-1-derived peptides to HLA-B*5101 was measured as previously described by using RMA-S-B*5101 cells[8].

Assays to determine recognition of peptide-pulsed or virus-infected targets. C1R and .221 cells expressing HLA-B*5101 or HLA-B*4801 were generated as previously described[30]. All cells were maintained in RPMI 1640 medium supplemented with 10% FCS and 0.15 mg ml^{-1} hygromycin B. Cytotoxicity of CD8$^+$ T cells for C1R-B*5101 cells pre-pulsed with peptide measured by the standard ^{51}Cr release assay was as previously described[8]. .221-B*4801 and .221 cells infected with NL4-3 or NL4-3 A146P mutant virus were used as target cells for intracellular cytokine staining assay.

Generation of the NL4-3 A146P mutant virus. The p82-2 plasmid containing the A146P mutation[4] was digested with BssHII and ApaI. The BssHII–ApaI 1.3-kb fragment was purified and then ligated into the same site of BssHII–ApaI-digested pNL-432 plasmid. To obtain pNL-432 including the A146P mutant (pNL-432 A146P), 293T cells were transfected with pNL-432 A146P using Lipofectamine 2000 (Invitrogen). Supernatants from transfected 293T cell cultures were stored at −80 °C.

Generation of CD8$^+$ T-cell clones and peptide-specific CD8$^+$ T-cell lines. Cytotoxic T lymphocyte (CTL) clones were generated from HIV-1-specific bulk-cultured T cells by limiting dilution as previously described[8]. Peptide-specific CD8$^+$ T-cell lines were generated by stimulating peripheral blood mononuclear cells (PBMCs) from the HLA-B*4801-positive HIV-1-seropositive individual KI-092 with the NI11 (NLQGQMVHQAI) peptide and then culturing them for 2 weeks[8]. Cytotoxicity of CD8$^+$ T cells for target cells pre-pulsed with peptide measured by the standard ^{51}Cr release assay was as previously described[8].

Suppression assay of HIV-1 replication by HIV-1-specific CTLs. The ability of HIV-1-specific CTLs to suppress HIV-1 replication was examined as previously described[30].

Intracellular cytokine staining assays. PBMCs from HIV-1-infected individuals were stimulated with the desired peptide (1 μM) and cultured for 12–14 days. These cultured PBMCs were assessed for IFN-γ-producing activity as previously described[30].

27. Tang, J. *et al.* Favorable and unfavorable HLA class I alleles and haplotypes in Zambians predominantly infected with clade C human immunodeficiency virus type 1. *J. Virol.* **76,** 8276–8284 (2002).
28. Agresti, A. & Coull, B. Approximate is better than 'exact' for interval estimation of binomial proportions. *Am. Stat.* **52,** 119–126 (1998).
29. Gatanaga, H., Hachiya, A., Kimura, S. & Oka, S. Mutations other than 103N in human immunodeficiency virus type 1 reverse transcriptase (RT) emerge from K103R polymorphism under non-nucleoside RT inhibitor pressure. *Virology* **344,** 354–362 (2006).
30. Tomiyama, H., Akari, H., Adachi, A. & Takiguchi, M. Different effects of Nef-mediated HLA class I down-regulation on HIV-1-specific CD8$^+$ T cell cytokine activity and cytokine production. *J. Virol.* **76,** 7535–7543 (2002).

LETTERS

An unexpected twist in viral capsid maturation

Ilya Gertsman[1,2], Lu Gan[1]†, Miklos Guttman[2], Kelly Lee[1], Jeffrey A. Speir[1], Robert L. Duda[3], Roger W. Hendrix[3], Elizabeth A. Komives[2] & John E. Johnson[1,2]

Lambda-like double-stranded (ds) DNA bacteriophage undergo massive conformational changes in their capsid shell during the packaging of their viral genomes. Capsid shells are complex organizations of hundreds of protein subunits that assemble into intricate quaternary complexes that ultimately are able to withstand over 50 atm of pressure during genome packaging[1]. The extensive integration between subunits in capsids requires the formation of an intermediate complex, termed a procapsid, from which individual subunits can undergo the necessary refolding and structural rearrangements needed to transition to the more stable capsid. Although various mature capsids have been characterized at atomic resolution, no such procapsid structure is available for a dsDNA virus or bacteriophage. Here we present a procapsid X-ray structure at 3.65 Å resolution, termed prohead II, of the lambda-like bacteriophage HK97, the mature capsid structure of which was previously solved to 3.44 Å (ref. 2). A comparison of the two largely different capsid forms has unveiled an unprecedented expansion mechanism that describes the transition. Crystallographic and hydrogen/deuterium exchange data presented here demonstrate that the subunit tertiary structures are significantly different between the two states, with twisting and bending motions occurring in both helical and β-sheet regions. We also identified subunit interactions at each three-fold axis of the capsid that are maintained throughout maturation. The interactions sustain capsid integrity during subunit refolding and provide a fixed hinge from which subunits undergo rotational and translational motions during maturation. Previously published calorimetric data of a closely related bacteriophage, P22, showed that capsid maturation was an exothermic process that resulted in a release of 90 kJ mol^{-1} of energy[3]. We propose that the major tertiary changes presented in this study reveal a structural basis for an exothermic maturation process probably present in many dsDNA bacteriophage and possibly viruses such as herpesvirus, which share the HK97 subunit fold[4].

HK97 is a favourable system for studying capsid maturation as capsid particles can be assembled in *Escherichia coli* from the expression of just two viral gene products, gp4 (protease) and gp5 (capsid subunit), and maturation can be triggered and analysed *in vitro* (Fig. 1) using chemical or low pH treatments[2,5–8] as opposed to the packaging of dsDNA, which induces maturation *in vivo*. During the maturation, subunit reorganization facilitates a particle expansion from 540 Å (prohead II) to 660 Å (head II) in diameter (Fig. 1). The kinetics of maturation were previously studied using time-resolved solution X-ray scattering[9] and the structures of the intermediates were determined with cryo-electron microscopy[6,7,10] and X-ray crystallography[2,5]. Near atomic resolution structures have characterized the late maturation states (balloon, head II), but only lower resolution cryo-electron microscopy models were previously available for the procapsid and expansion intermediate (EI) forms, which used the 3.44 Å head II structure[11] as a basis for pseudo atomic models. The previous 12 Å resolution cryo-electron microscopy study of prohead II suggested that most of the capsid structural changes in expansion were the result of rigid-body rotations and translations of the central domains of the subunit, whereas the E-loop and N-arm regions moved independently[6].

Here we report a 3.65 Å resolution X-ray crystal structure of W336F, E-loop truncated prohead II (Protein Data Bank accession number 3E8K) that changes the previous conceptions of capsid maturation (crystallographic statistics listed in Supplementary Table 1). These mutations did not affect the assembly of the capsid or its ability to undergo maturation. The structure reveals that three-fold contacts between subunits, mediated by 'P-loops' as well as their surrounding β-strands on each subunit, are preserved during maturation from prohead II to head II. As a result, the previously proposed rigid subunit motions at lower resolution could now be resolved as domain motions corresponding to a twist of the subunit about three β-strands (βD, βJ and βI) (Fig. 1b), and a simultaneous bending and unwinding of the long (spine) helix with respect to the fixed three-fold interaction sites. The extent of the subunit twist and the helix bend vary among subunits and depend on their quasi-equivalent position.

The overall morphologies between the prohead II and head II states are very distinct. The subunits in prohead II are oriented radially relative to the capsid surface, but are roughly tangential in head II (Fig. 1c–e). A notable feature of prohead II, which was seen in the previous cryo-electron microscopy study, is that the skewed hexamers comprising trimers of subunits with a trapezoidal arrangement give the hexamers a pseudo two-fold appearance.

The refined P-loop contacts in prohead II bear a striking similarity to the same contacts in head II. The P-loop of each subunit is tightly associated with the P-loop of two other subunits from separate capsomers at all three-fold and quasi-three-fold axes (Fig. 2). In the previous cryo-electron-microscopy-based model, the P-loop of prohead II was kept fixed relative to the subunit core, changing the trimer associations when compared with those in head II. It is now clear that the position and quaternary interactions of the P-loops and surrounding β-strands (region coloured blue in Fig. 2c) are unchanged during expansion, demonstrating that it functions as a fixed point of subunit interaction in an otherwise highly plastic quaternary structure. Figure 2b, c shows salt-bridge interactions between Glu 344 and Glu 363 from two of the β-strands surrounding the P-loop on one subunit with Arg 194 (located on the turn following the spine helix) and Arg 347 (located on the P-loop) of a neighbouring threefold-related subunit. The salt bridges as well as a putative metal-binding site coordinating three glutamate (E348) residues directly underneath each three-fold axis (Fig. 2b) remain unchanged during capsid maturation. Three of these residues (R194, E344 and E363) are proximal to the borders of the region that remains fixed during maturation, defining the boundaries of the pivot points of

[1]Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037, USA. [2]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California 92037, USA. [3]Pittsburgh Bacteriophage Institute & Department of Biological Sciences, University of Pittsburgh, Pennsylvania 15260, USA. †Present address: Division of Biology, California Institute of Technology, Pasadena, California 91125, USA.

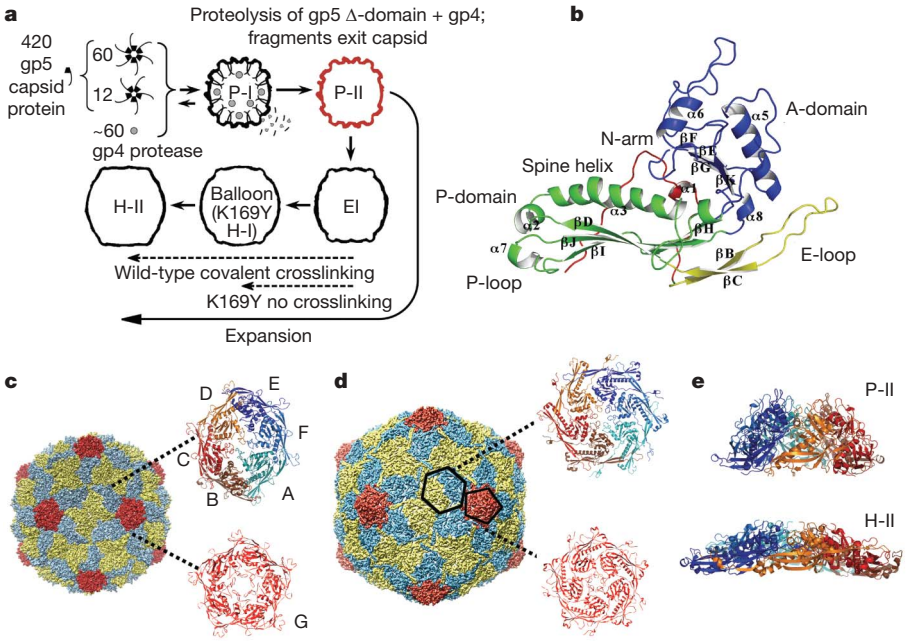


Figure 1 | HK97 assembly and morphology. a, The 384-residue gp5 subunit initially assembles into hexameric and pentameric oligomers, termed capsomers, that first assemble to form the prohead I capsid (P-I). The $T = 7$ *laevo* particle is composed of 12 pentamers and 60 hexamers and encapsidates approximately 60 copies of gp4 protease[23–25]. Expression with a defective protease produces a prohead I particle that can be disassembled *in vitro* into free capsomers and re-assembled when exposed to specific chemical treatments[22]. When active gp4 is present, particles spontaneously mature to the 13-MDa prohead II (P-II) form after digestion of residues 2–103 from all subunits. Crosslinking occurs in the wild-type particle after formation of the EI state. Crosslinks (isopeptide bond) form between Lys 169 and Asn 356 located on different subunits. A crosslink-defective mutant, K169Y, expands to head I, a state nearly identical to balloon minus the crosslinks. Wild-type balloon undergoes a final expansion step to head II in which the pentons become more protruded and form one last class of crosslinks, with a molecular topology similar to chainmail[2,26]. **b**, Crystal structure of subunit D of prohead II at 3.65 Å. **c**, 3.65 Å electron density map (displayed as a solid surface) of the full prohead II capsid, contoured at ~1σ in Chimera. The prohead II hexamers and pentamers are shown alongside the capsid with the seven subunits of the viral asymmetric subunit labelled A–F for the hexamers and G for the pentamers. **d**, A calculated electron density map of the head II capsid shown at 3.65 Å, also rendered at ~1σ. **e**, Prohead II and head II hexamers shown tangential to the capsid surface (rotated 90° from view **c** and **d**).

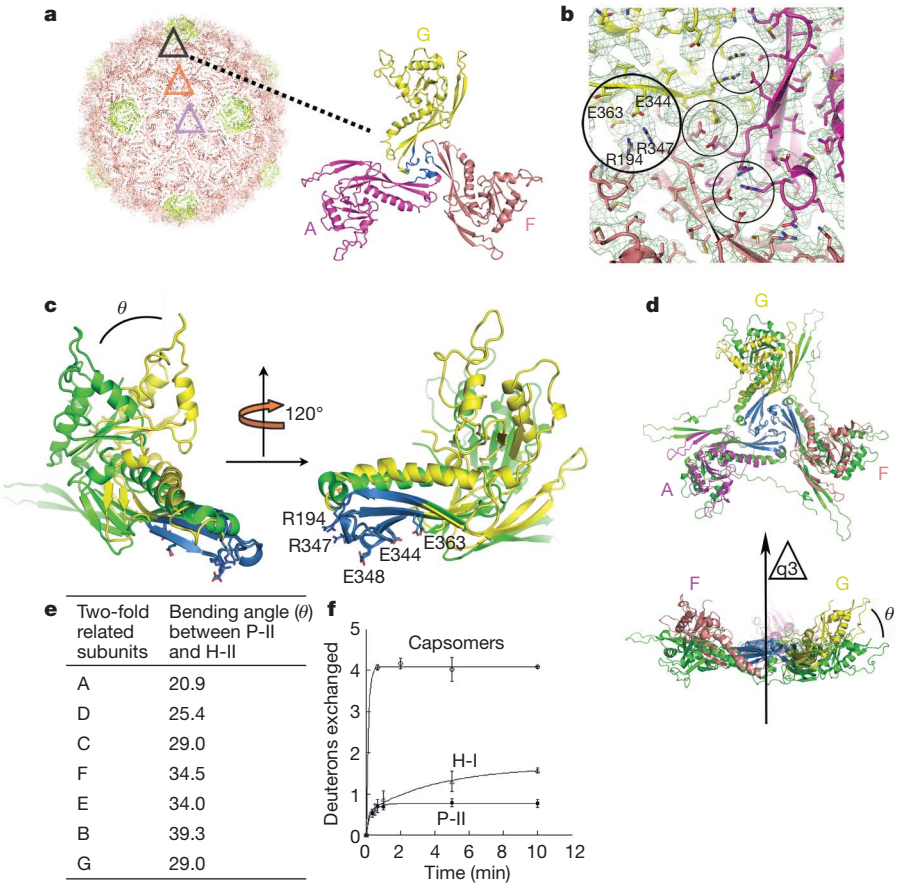
tertiary rearrangement (Fig. 2c). In accord with this newly recognized structural constraint, the tertiary structure of the subunit is now seen to have a significant twist about the P-domain β-sheet (Supplementary Movie 1).

To corroborate the conclusions from the crystallographic data, we characterized the dynamics of the three-fold P-loop interactions with H/^{2}H exchange coupled to matrix-assisted laser desorption/ionization (MALDI) mass spectrometry on prohead II, head I and free capsomers. The technique measures the solvent accessibility of amide protons (in native proteins in solution) whose rate of exchange with deuterium is influenced by secondary, tertiary and quaternary structure interactions[12,13]. After incubation in deuterium, the capsid protein is digested with pepsin protease and the masses of previously determined peptide fragments are quantified. Regions with greater solvent accessibilities will have larger shifts in their mass envelopes, which are quantified as described in the Methods. A K169Y mutant was used instead of wild-type head II for the study because covalent crosslinks inhibited efficient pepsin digestion and subsequent analysis by mass spectrometry. The mutant is able to expand through similar intermediate forms as wild-type prohead II (Fig. 1a), although maturation stops at the penultimate, head I state, which was shown by crystallography to have very similar subunit tertiary structures compared to wild-type head II[5]. The crosslink-defective mutant therefore permitted comparisons of H/^{2}H exchange profiles between subunits in prohead II and subunits in a virtually mature particle form. H/^{2}H exchange was also performed on capsomers that were disassociated from the prohead I state and were no longer able to form three-fold P-loop associations. One of the peptide fragments spanned residues 345–353 of the P-loop (coloured lime-green in Fig. 2a), which lies at the junction of the trimer interface. As seen in Fig. 2e, this P-loop fragment is highly solvent protected in both the prohead II and head I states, whereas in free capsomers it is nearly five times more solvent accessible. Quaternary interactions are therefore limiting the rate of amide proton exchange in these intact particle forms, whereas P-loops in the unassociated capsomers are more free to exchange. Data generated for EI (Supplementary Fig. 1) yielded nearly identical exchange profiles as seen for prohead II and head I, verifying the presence of P-loop interactions during intermediate stages of expansion as well. Consistent with the prohead II crystal structure, the H/^{2}H data confirmed that strong interactions remained fixed at the P-loop three-fold sites, despite the large subunit rotational motions.

The magnitudes of rotation that bring the prohead II subunit into the head II conformation were measured (Fig. 2e) by superimposing the residues behind the fixed region coloured blue in Fig. 2c. The measurements therefore directly relate to the degree of tertiary twisting, which at lower resolution was quantified as whole-subunit rotations in previous studies[6]. Subunits closest to the pseudo two-fold axis (A and D) undergo the least rotation, whereas those farthest from it (B and E) undergo the most rotation.

One of the fixed anchor points, Arg 194, resides several residues amino-terminal to the spine helix. Most of the subunit beyond the fixed P-domain region (coloured blue in Fig. 2c) twists as a rigid unit, causing significant bending of the helix, which is fixed at its N-terminal end. The degree of helix bending is therefore proportional to the extent of β-strand twisting (Supplementary Movie 1). The helix deformation in prohead II can be seen in Fig. 3a and Supplementary Movie 3. Subunits B, C, E, F and G show marked helix bending whereas subunits A and D show straighter helices as well as smaller twisting motions in the P-domain β-sheet.

To examine the dynamics of the spine helix in solution, H/^{2}H exchange was measured for a peptide spanning residues 206–216,



Two-fold related subunits	Bending angle (θ) between P-II and H-II
A	20.9
D	25.4
C	29.0
F	34.5
E	34.0
B	39.3
G	29.0

Figure 2 | P-loops located at three-fold axes act as invariant pivot points. **a**, Ribbon representation of prohead II. The orange triangle represents an icosahedral three-fold axis; black and magenta triangles represent two quasi-three-fold positions. A magnified view of subunits at a quasi-three-fold axis is shown viewed from outside of the capsid. Residues 345–353 of the P-loop are coloured lime-green, and represent the peptide fragment analysed by H/^{2}H exchange. **b**, Side-chain interactions at three-fold axes that remain invariant during expansion, viewed from the interior of the capsid directly underneath a quasi-three-fold axis, 180° from the view of the trimer in **a**. Electron density is contoured at 1σ. The three outer circles highlight salt bridges whereas the centre circle highlights three glutamates (E348) coordinated at a putative metal-binding site. **c**, Subunit G of prohead II (yellow) and head II (green) have been aligned by the region of the P-domain which remains invariant (blue). (This motion is best captured in Supplementary Movie 1.) **d**, A prohead II trimer (subunits A, F, G as shown in **a**) is aligned on the head II trimer (green) by the regions that remain invariant (blue), illustrating the rotational motions in respect to the fixed trimeric interactions. The upper panel is looking down a quasi-three-fold axis (q3) whereas the lower panel shows a perpendicular view, tangential to the capsid surface. **e**, Table shows the angles of rotation (θ) of each subunit from prohead II to head II as illustrated in both **c** and **d**. **f**, H/^{2}H exchange curve of a peptide fragment spanning residues 345–353 of the P-loop (coloured lime-green in **a**) shown for prohead II, head I and free capsomer states. Time points are taken from 30 s to 10 min, with error bars representing standard deviations from the average of three independent experiments, with 2–3 measurements per experiment (6–9 total measurements for each time point).

which covers the bent region. The average amount of deuterium exchanged in this region of prohead II is nearly five times greater than in head I, showing a more canonical helical structure with stronger hydrogen bonding in the mature head I form (Fig. 3b). H/^{2}H measurements of the first expansion intermediate, EI, were also performed. The helix peptide shows nearly identical solvent exchange for EI as head I, indicating that the increased hydrogen bonding in the helix occurs in the initial stage of expansion. The helix in the free capsomer state shows a similar level of solvent accessibility as prohead II, indicating that the helix distortion is not just a result of the quaternary arrangement enforced in the intact capsid, but is probably occurring at the level of capsomer assembly and facilitated by interactions of the Δ-domain (residues 2–103 that function as a scaffold and are cleaved off of prohead I to form prohead II).

Quaternary associations probably induce different degrees of strain in the local tertiary structures of the seven quasi-equivalent subunits. The subunits are not only in a skewed arrangement in the prohead II hexamer, but they also show different orientations depending on their positions in the hexamer. While the long axes of subunits B, C, E and F lie more radial to the capsid surface, subunits A and D lie more parallel to the capsid surface and therefore do not need to rotate as much to assume their orientations in the mature hexamer (Supplementary Movie 2). Because P-loop contacts are preserved during maturation, there is a strong correlation between the orientation of the subunit relative to the capsid surface and the change in tertiary structure between prohead II and head II, with the more tangential subunits showing less tertiary structure change and the more radial displaying the larger tertiary structure change.

The combined crystallographic and H/^{2}H exchange data demonstrate that the large subunit rotations concomitant with expansion from prohead II to head II are facilitated by a tertiary structural transition—the twist of the subunit core about a fixed hinge. H/^{2}H exchange data of the helix, which appears to be bent in concert with the overall hinging motions, indicates that most of the change in tertiary structure occurs during the initial and irreversible expansion from prohead II to EI[8,14] (Fig. 3c). This is reasonable considering nearly 60% of the expansion in size occurs in the first transition, as well as the symmetrization of the hexamers[8] (Supplementary Movie 2). We propose that the bent helix and twisted β-strand in prohead II place the subunits in a strained conformation of elevated free energy

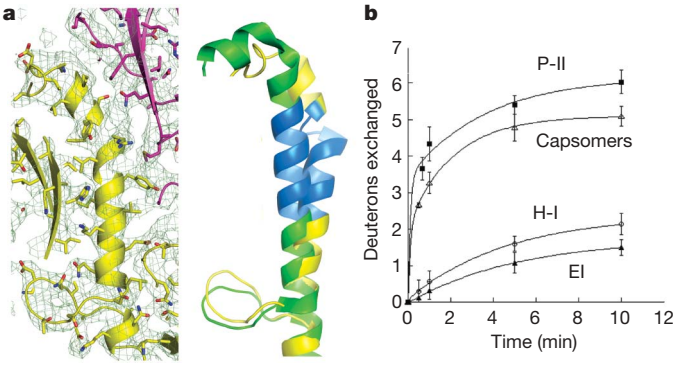


Figure 3 | Spine helix bends during maturation. a, The spine helix (yellow) is shown for subunit F of prohead II both in its corresponding electron density on the left (1σ), and aligned with head II (green) on the right. The subunits from the two states were aligned using the subunit core that acts mostly as a rigid body (residues 230–383) with an r.m.s.d of 1.3 Å or better for each alignment. The region coloured blue represents the fragment spanning residues 206–216, analysed by H/^{2}H exchange. **b**, H/^{2}H exchange rate curves comparing deuterium exchange between prohead II, EI I, head I and free capsomer helix fragment.

and that this accounts for both the meta-stability of prohead II and the driving force for the initial expansion to EI. The three-fold interactions at the P-loops stabilize inter-capsomer interactions during the expansion. Capsid integrity is augmented after transition to EI, which is competent for covalent crosslinking in the three-fold region. The energy sources for the distorted tertiary structure in prohead II probably stem from the initial assembly, in which the Δ-domains (residues 2–103) of each subunit putatively act as molecular scaffolds that promote capsomer assembly. The favourable association of Δ-domains in this early assembly product may induce the strained conformation (Fig. 4). The high level of deuterium exchange observed in the spine helix of free capsomers supports our hypothesis that the bent subunit conformation exists at the stage of capsomers, not just fully assembled capsid. Δ-domains interact in a trimer arrangement in the hexamers of

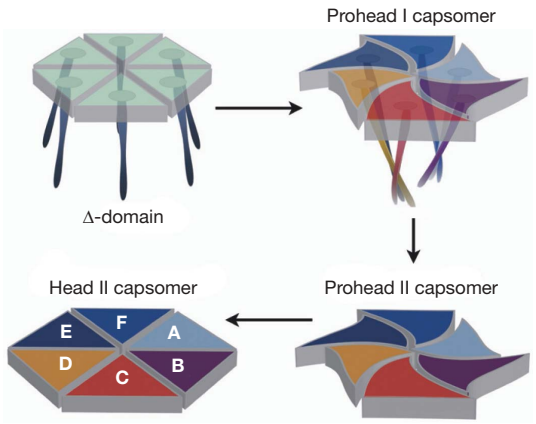


Figure 4 | A working hypothesis for the formation, meta-stability and subsequent maturation of HK97, represented with a single hexamer. Individual subunits are first assembled into hexamers and pentamers (the top-left panel is a hypothetical representation of the initial subunit organization). Based on H/^{2}H exchange data, subunit tertiary structures are distorted in free capsomers and we propose that the hexamers are skewed (top-right panel). Prohead I is formed by assembly of hexamers and pentamers into a $T = 7$ particle with Δ-domains attached. After proteolysis of the Δ-domains to form prohead II, the skewed hexamers and distorted tertiary structures are preserved by quaternary structure interactions in the particle, raising the free energy of the particle to a meta-stable state maintained in a local minimum. Perturbation of these particles by dsDNA packaging (*in vivo*) or lowering the pH (*in vitro*) lowers the energy barrier, leading to an exothermic expansion of the particles producing symmetric hexamers and undistorted subunit tertiary structures.

prohead I[15], which assume a skewed symmetry similar to prohead II. Although prohead I, prepared without the viral protease, is resistant to expansion when exposed to conditions that expand prohead II, cryo-electron microscopy of prohead I particles heated to 55 °C showed a reversible transition to an EI-like state[15]. That study showed that heating the prohead I particles causes a disruption of Δ-domain interactions, enabling the particle to expand beyond the prohead II state to a state in which the hexamers were symmetrical. Based on these data we argue that upon disruption of Δ-domain interactions and the formation of symmetric hexamers, the tertiary strain in the subunits is relieved. When the Δ-domains are present, as they are in prohead I, cooling the particles causes the Δ-domains to re-associate, which induces the skewed hexamers and strained tertiary structure. When the Δ-domain is absent as in prohead II, the particle is trapped in an elevated local energy minimum until it is perturbed to expand either by DNA packaging *in vivo*, or by chemical perturbation *in vitro*.

One study[16] previously proposed a mechanism for P22 expansion where the procapsid subunit exists as a late-folding intermediate that undergoes further tertiary changes en route to the lower energy, mature conformation. Such a mechanism is now evident in HK97, and may be the driving force for expansion. Systems in which tertiary structure folding events, comparable to those presented here for HK97, have been characterized include the CA domain of the HIV capsid protein, which has been shown to require a kinking of a helix to induce dimer activation[17]. Although such tertiary structural changes have not been characterized at high resolution in other dsDNA bacteriophage and viruses, they may be present in P22, T4, T7, φ29 and possibly animal viruses such as herpesviruses, which all share an HK97-like fold[4,18–21].

METHODS SUMMARY

Mutagenesis and crystallography. The W336F mutation suppresses the spontaneous expansion observed in wild-type proheads and therefore increased the homogeneity of prohead II preparations. The E-loop was truncated between residues 159–171 to improve crystallization, as previous studies showed that the tip of the full-length E-loop was partially disordered and protruded from the capsid surface[6]. Crystals were grown using the hanging-drop vapour diffusion method with a mother liquor consisting of 0.1 M CHES buffer, pH 9.0, 200 mM manganese chloride and 2.3–3.0% Peg 4000. The addition of CHES and Peg to the manganese chloride caused precipitation of much of the manganese. Only the supernatant from the precipitated solution was used for crystallization. A 200 mM final concentration of NDSB-211 (Hampton Research) was added to the drop. An atomic model for the prohead II structure was initially derived by rigid-body fitting of the refined 3.44 Å structure of the mature head II coordinates (Protein Data Bank 1OHG) into the prohead II electron density. The initial phases for molecular replacement were derived from the previously solved 12 Å cryo-electron microscopy structure of prohead II.

H/^{2}H exchange and sample preparation. H/^{2}H exchanged samples were analysed on a DE-STR MALDI-TOF mass spectrometer. Concentrated HK97 capsids (~40 mg ml^{-1} protein concentration) were diluted nearly sevenfold in a final D_2O concentration of 85%, buffered with 20 mM Tris, pH 7.5 and containing 200 mM sodium chloride. Preparation of capsomers and the late expansion intermediate, head I, used in the H/^{2}H exchange study was performed and monitored as previously described[5,22]. The EI-I particle form was obtained by treatment of prohead II with 10% isobutanol followed by a 15-min incubation. Isobutanol is one of many chemical conditions that has been previously shown to cause capsid maturation *in vitro*[23], and was used for its compatibility with analysis by MALDI mass spectrometry.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 September; accepted 8 December 2008.
Published online 8 February 2009.

1. Smith, D. E. *et al.* The bacteriophage straight φ29 portal motor can package DNA against a large internal force. *Nature* **413,** 748–752 (2001).
2. Wikoff, W. R. *et al.* Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* **289,** 2129–2133 (2000).
3. Steven, A. C., Heymann, J. B., Cheng, N., Trus, B. L. & Conway, J. F. Virus maturation: dynamics and mechanism of a stabilizing structural transition that leads to infectivity. *Curr. Opin. Struct. Biol.* **15,** 227–236 (2005).

4. Baker, M. L., Jiang, W., Rixon, F. J. & Chiu, W. Common ancestry of herpesviruses and tailed DNA bacteriophages. *J. Virol.* **79,** 14967–14970 (2005).
5. Gan, L. *et al.* Capsid conformational sampling in HK97 maturation visualized by X-ray crystallography and cryo-EM. *Structure* **14,** 1655–1665 (2006).
6. Conway, J. F. *et al.* Virus maturation involving large subunit rotations and local refolding. *Science* **292,** 744–748 (2001).
7. Lata, R. *et al.* Maturation dynamics of a viral capsid: visualization of transitional intermediate states. *Cell* **100,** 253–263 (2000).
8. Lee, K. K. *et al.* Virus capsid expansion driven by the capture of mobile surface loops. *Structure* **16,** 1491–1502 (2008).
9. Lee, K. K., Tsuruta, H., Hendrix, R. W., Duda, R. L. & Johnson, J. E. Cooperative reorganization of a 420 subunit virus capsid. *J. Mol. Biol.* **352,** 723–735 (2005).
10. Wikoff, W. R. *et al.* Time-resolved molecular dynamics of bacteriophage HK97 capsid maturation interpreted by electron cryo-microscopy and X-ray crystallography. *J. Struct. Biol.* **153,** 300–306 (2006).
11. Helgstrand, C. *et al.* The refined structure of a protein catenane: the HK97 bacteriophage capsid at 3.44 Å resolution. *J. Mol. Biol.* **334,** 885–899 (2003).
12. Mandell, J. G., Baerga-Ortiz, A., Akashi, S., Takio, K. & Komives, E. A. Solvent accessibility of the thrombin-thrombomodulin interface. *J. Mol. Biol.* **306,** 575–589 (2001).
13. Croy, C. H., Bergqvist, S., Huxford, T., Ghosh, G. & Komives, E. A. Biophysical characterization of the free IκBα ankyrin repeat domain in solution. *Protein Sci.* **13,** 1767–1777 (2004).
14. Lee, K. K. *et al.* Evidence that a local refolding event triggers maturation of HK97 bacteriophage capsid. *J. Mol. Biol.* **340,** 419–433 (2004).
15. Conway, J. F. *et al.* A thermally induced phase transition in a viral capsid transforms the hexamers, leaving the pentamers unchanged. *J. Struct. Biol.* **158,** 224–232 (2007).
16. Tuma, R., Prevelige, P. E. Jr & Thomas, G. J. Jr. Mechanism of capsid maturation in a double-stranded DNA virus. *Proc. Natl Acad. Sci. USA* **95,** 9885–9890 (1998).
17. Ivanov, D. *et al.* Domain-swapped dimerization of the HIV-1 capsid C-terminal domain. *Proc. Natl Acad. Sci. USA* **104,** 4353–4358 (2007).
18. Jiang, W. *et al.* Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolutions. *Nature Struct. Biol.* **10,** 131–135 (2003).
19. Fokine, A. *et al.* Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc. Natl Acad. Sci. USA* **102,** 7163–7168 (2005).
20. Agirrezabala, X. *et al.* Quasi-atomic model of bacteriophage t7 procapsid shell: insights into the structure and evolution of a basic fold. *Structure* **15,** 461–472 (2007).
21. Morais, M. C. *et al.* Conservation of the capsid structure in tailed dsDNA bacteriophages: the pseudoatomic structure of φ29. *Mol. Cell* **18,** 149–159 (2005).
22. Xie, Z. & Hendrix, R. W. Assembly *in vitro* of bacteriophage HK97 proheads. *J. Mol. Biol.* **253,** 74–85 (1995).
23. Duda, R. L. *et al.* Structural transitions during bacteriophage HK97 head assembly. *J. Mol. Biol.* **247,** 618–635 (1995).
24. Duda, R. L., Martincic, K. & Hendrix, R. W. Genetic basis of bacteriophage HK97 prohead assembly. *J. Mol. Biol.* **247,** 636–647 (1995).
25. Conway, J. F., Duda, R. L., Cheng, N., Hendrix, R. W. & Steven, A. C. Proteolytic and conformational control of virus capsid maturation: the bacteriophage HK97 system. *J. Mol. Biol.* **253,** 86–99 (1995).
26. Duda, R. L. Protein chainmail: catenated protein in viral capsids. *Cell* **94,** 55–60 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank V. Reddy for assistance with crystallographic studies and for discussions. We thank R. Huang for providing HK97 capsomer samples and for discussions, and T. Matsui for help with X-ray data collection. We thank B. Firek and C. Moyer for mutagenesis of the HK97 constructs used in the study. We also thank B. Szymczyma for material used in the study. We also thank I. Wilson for discussions. We thank the staffs at beamlines 14-BMC and 23-ID-D of the Advanced Photon Source for assistance in data collection. This work was supported by NIH grants RO1 AI40101 (to J.E.J), RO1 GM47795 (to R.W.H) and NIH Training Grant GM08326.

Author Contributions I.G. was the lead investigator that crystallized the prohead II particles, collected the X-ray data and determined and refined the structure. He also collected and interpreted the hydrogen/deuterium exchange data and prepared the first draft of the paper. L.G. helped with the initial crystallography of the prohead II particles. M.G. helped with the initial collection and interpretation of the hydrogen/deuterium exchange data. K.L. characterized the kinetics and parameters associated with the prohead II to EI transition facilitating the hydrogen/deuterium exchange studies of the EI intermediate. J.A.S. made important contributions during the refinement of the prohead II structure. R.L.D. and R.W.H. developed the HK97 expression system that allowed the studies to be performed, prepared the prohead II mutations that facilitated the production of crystals that diffracted to high resolution, contributed valuable advice for handling the particles and helped in writing the manuscript. E.A.K. supervised the hydrogen/deuterium exchange studies that were all performed in her laboratory. J.E.J. supervised the crystallography aspect of the project, coordinated the overall project and helped in writing the manuscript.

Author Information The sequence for W336F, E-loop truncated prohead II has been deposited in the Protein Data Bank under accession number 3E8K. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.E.J. (jackj@scripps.edu).

METHODS

Mutagenesis. The W336F point mutation was generated from wild-type plasmid by site-directed mutagenesis. A truncation mutation was made in the tip of the E-loop of each subunit, a region seen to be dynamic in prohead II in both cryo-electron microscopy and previous crystallographic studies. Splicing by overlap extension was performed on the W336F construct, replacing residues 159–171 with residues APGD, a sequence known to promote formation of a reverse turn. The result was a truncated loop that was fully visible in the electron density of the crystal structure. Expression of gp5 capsid protein and gp4 protease was completed in an *E. coli* T7 expression system and purification of HK97 prohead II was performed as previously described[27]. W336F mutant assembles into an intact capsid with similar efficiency as wild type. W336F prohead II is able to crosslink and expand under acidic conditions, but at a slower rate than wild-type prohead II (R.L.D. and R.W.H., unpublished data).

Crystallographic processing. Crystallographic data were collected at the Advanced Photon Source synchrotron at Argonne National Laboratories, beamlines 23-IDD and 14-BMC. The room-temperature diffraction data from 29 crystals was indexed, integrated and scaled using the HKL2000 suite[28]. The crystals belong to the space group *I*222. Reflections with an $I/\sigma(I)$ of less than 0 were discarded during scaling. Partial reflections were then scaled up to whole reflections using CCP4's SCALA program, followed by scaling of all reflections also with SCALA. The orientation of the particle was confirmed with a five-fold self-rotation function using the GLRF program[29] to determine which of two possible particle orientations is maintained in the unit cell. Averaging and phase extension was done with CCP4 and RAVE using a previously determined 12 Å cryo-electron microscopy model[6] to build the mask used for molecular replacement. The head II structure was manually fit into the prohead II density followed by rigid body refinement in CNS. The majority of each subunit fit well into the prohead II map, but certain regions in the A-domain loops, P-domain and E-loop required significant additional adjustments to improve the fit to the experimental density. These regions were manually fit in 'Coot'[30] followed by simulated annealing in real space using Rsref2000 (ref. 31). Energy minimization and geometry refinement of the rebuilt residues was then done in CNS. The regions that required conformation refitting included residues 289–294 of subunits A and D, 298–305 of A–G, and residues 193–215 of the spine helix. Regions that were hinged include the E-loop, N-arm and P-domain β-sheets. Residues 104–118 of the N-arm were disordered with no visible electron density. The first residue of the N-arm for which electron density could be seen varied for each subunit. The most N-arm density was seen for subunit A, which was visible starting from residue 119, whereas density for the other subunits started between residues 120 and 127. Electron density for the tip of the E-loops, residues 158–162, also appears weakly, probably due to conformational flexibility in this region.

Bending angles between subunits from prohead II and head II states (Fig. 2) were calculated by deconvoluting matrices needed to align coordinates representing the refined prohead II structure with that of head II structure fit into the prohead II map. The two particle states were initially least squares aligned by the P-loops (346–357), which remained fixed during expansion. r.m.s.d. values ranged from 0.67 Å to 1 Å. Matrices were than calculated for the alignment of the subunit cores (residues 230–383) of prohead II and head II from the initial P-loop aligned state. r.m.s.d. values for these least squares alignments ranged from 1.1 Å to 1.3 Å. Residues in this core region remain mostly rigid during expansion and are all located C-terminal to the E-loop. Residues N-terminal to the E-loop were not used for alignment, as these residues are all involved in major structural movements between the two states.

H/^{2}H exchange. Samples were incubated in an 85% D_2O solution at pH 7.5 for various time periods, then quenched by the addition of a pH 2.5 non-deuterated quench solution containing trifluoracetic acid (final D_2O concentration of 9.0%). After quench, all samples were kept on ice in a 4 °C fridge. Protein was digested with 50 µl pepsin-coated beads (Pierce) for 5 min. Beads were removed by centrifugation, whereas supernatant was flash frozen in liquid N_2. Samples were thawed individually, mixed 1:1 with alpha C matrix and vacuum crystallized on a Maldi plate. H/^{2}H exchanged samples were analysed on a DE-STR MALDI-TOF mass spectrometer. The total number of deuterons exchanged was calculated by subtracting the centroid of the mass envelope from the non-deuterated control from the centroid of the deuterated mass envelopes. The error (standard deviation) was estimated from the average of three independent experiments with 2–3 measurements recorded for each experiment (total of 6–9 measurements for each time point). Back exchange was calculated as 42% using a peptide in the N-arm region (residues 117–126), which exchanged amide protons for deuterium completely within 20 s. A separate control for back exchange was performed using an 11-residue unstructured synthetic polypeptide, which showed similar back exchange values as the N-arm fragment (117–126). N-terminal amide protons of peptide fragments were not considered exchangeable residues as they are not expected to retain deuterium after quenching, nor were proline residues.

Deuterium incubations were performed for up to 10 min, enabling measurement of amide protons exchanging at both a fast ($>1\ \mathrm{min}^{-1}$) and intermediate rate (0.01 to $1\ \mathrm{min}^{-1}$)[32]. Longer incubations measuring slow exchange rates ($<0.01\ \mathrm{min}^{-1}$) were not done in this study. Amide protons exchanging at intermediate to slow rates are generally a result of solvent protection due to either secondary structure or protein–protein interactions. The measured data for all fragments was best fit to either a single or two-exponential model accounting for deuterons exchanging at only a fast rate, or both a fast and intermediate rate respectively. The following equation represents the two-exponential fit:

$$D = N_{\mathrm{fast}}(1 - e^{-k_{\mathrm{fast}}t}) + N_{\mathrm{inter}}(1 - e^{-k_{\mathrm{inter}}t})$$

where D is the total number of deuterons exchanged at time t, N_{fast} is the number of deuterons exchanging at a fast rate, k_{fast}, and N_{inter} is the number of deuterons exchanging at an intermediate rate, k_{inter}. The fast exchanging amide protons had nearly all exchanged by the first time point, so k_{fast} was estimated as described previously[12]. All fragments were identified with tandem mass spectrometry (MS/MS) using a Q-star mass spectrometer.

27. Duda, R. L. Protein chainmail: catenated protein in viral capsids. *Cell* **94,** 55–60 (1998).
28. Otwiniowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
29. Tong, L. R. & Rossmann, M. G. The locked rotation function. *Acta Crystallogr. A* **46,** 783–792 (1990).
30. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D.* **60,** 2126–2132 (2004).
31. Korostelev, A., Bertram, R. & Chapman, M. S. Simulated-annealing real-space refinement as a tool in model building. *Acta Crystallogr. D* **58,** 761–767 (2002).
32. Kang, S. & Prevelige, P. E. Jr. Domain study of bacteriophage p22 coat protein and characterization of the capsid lattice transformation by hydrogen/deuterium exchange. *J. Mol. Biol.* **347,** 935–948 (2005).

LETTERS

FGF signalling during embryo development regulates cilia length in diverse epithelia

Judith M. Neugebauer[1], Jeffrey D. Amack[1]†, Annita G. Peterson[1], Brent W. Bisgrove[1] & H. Joseph Yost[1]

Cilia are cell surface organelles found on most epithelia in vertebrates. Specialized groups of cilia have critical roles in embryonic development, including left–right axis formation. Recently, cilia have been implicated as recipients of cell–cell signalling[1,2]. However, little is known about cell–cell signalling pathways that control the length of cilia[3]. Here we provide several lines of evidence showing that fibroblast growth factor (FGF) signalling regulates cilia length and function in diverse epithelia during zebrafish and *Xenopus* development. Morpholino knockdown of FGF receptor 1 (Fgfr1) in zebrafish cell-autonomously reduces cilia length in Kupffer's vesicle and perturbs directional fluid flow required for left–right patterning of the embryo. Expression of a dominant-negative FGF receptor (DN-Fgfr1), treatment with SU5402 (a pharmacological inhibitor of FGF signalling) or genetic and morpholino reduction of redundant FGF ligands Fgf8 and Fgf24 reproduces this cilia length phenotype. Knockdown of Fgfr1 also results in shorter tethering cilia in the otic vesicle and shorter motile cilia in the pronephric ducts. In *Xenopus*, expression of a dn-fgfr1 results in shorter monocilia in the gastrocoel roof plate that control left–right patterning[4] and in shorter multicilia in external mucociliary epithelium. Together, these results indicate a fundamental and highly conserved role for FGF signalling in the regulation of cilia length in multiple tissues. Abrogation of Fgfr1 signalling downregulates expression of two ciliogenic transcription factors, *foxj1* and *rfx2*, and of the intraflagellar transport gene *ift88* (also known as *polaris*), indicating that FGF signalling mediates cilia length through an Fgf8/Fgf24–Fgfr1–intraflagellar transport pathway. We propose that a subset of developmental defects and diseases ascribed to FGF signalling are due in part to loss of cilia function.

FGF ligands bind and activate cell surface FGFRs to mediate multiple processes during embryogenesis. One ligand, Fgf8, has been proposed to have divergent roles in left–right patterning[5–9], as a left determinant in mouse and a right determinant in chick and rabbit. Experimental manipulations of FGFR function allow cell-autonomous alterations of FGF signalling not possible with manipulations of multiple secreted ligands that activate a given receptor. Using this approach, we investigated the roles of Fgfr1 in zebrafish development. To elucidate the role of Fgfr1 signalling in left–right development, we analysed the expression of *southpaw* (*spaw*; the zebrafish homologue of mouse *Nodal*), the earliest known asymmetrically expressed gene[10]. Knockdown of Fgfr1 with two distinct antisense morpholinos (MOs) perturbed the normal left-sided expression of *spaw* in the lateral plate mesoderm (Fig. 1a–c). Ets transcription factors *pea3* and *erm*, downstream targets of FGF signalling[11], were downregulated in *fgfr1* morphants (Supplementary Fig. 1), indicating the efficacy of MO knockdown. Markers of notochord (*no tail* (*ntl*), *lefty1*, *sonic hedgehog*)[12,13] and floorplate (*sonic hedgehog*) were found to be normal in *fgfr1* morphants (Supplementary Fig. 2), indicating that the barrier role of the embryonic midline is intact. These results indicate Fgfr1 signalling is required early in left–right development, preceding asymmetric expression of *spaw*.

spaw asymmetry is dependent on Kupffer's vesicle (KV), a ciliated epithelium structure that creates directional fluid flow[12–14], analogous to 'nodal flow' in mouse[15]. *fgfr1* messenger RNA is expressed in KV and surrounding tailbud (Fig. 1d, e). To determine whether FGF signalling functions cell-autonomously in KV cells to control *spaw* asymmetry, we generated chimaeric DFC$^{fgfr1\ MO\text{-}1}$ embryos in which *fgfr1* is knocked down in DFC/KV (dorsal forerunner cells; KV precursor cells) lineages[12] but not in the rest of the embryo. Similar to embryo-wide knockdown of *fgfr1*, DFC$^{fgfr1\ MO\text{-}1}$ embryos had significant alterations in *spaw* expression relative to DFC$^{control\ MO}$ ($P < 1.19 \times 10^{-5}$; Fig. 1c, f–h). As an important control, the effects

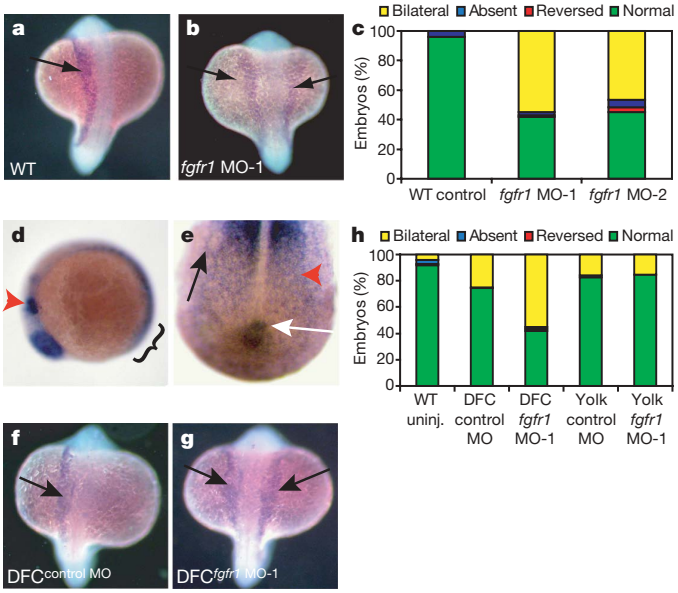


Figure 1 | Cell autonomous FGF signalling in Kupffer's vesicle controls left–right patterning. a, **b**, Dorsal view of left-sided *spaw* expression (arrow) in wild type (WT) (**a**), and bilateral expression in *fgfr1* MO-1 18–20-somite-stage embryos (**b**). **c**, Percentages of normal (left-sided), reversed, bilateral and absent *spaw* in WT control ($n = 99$), *fgfr1* MO-1 ($n = 117$) and *fgfr1* MO-2 ($n = 120$). **d**, **e**, *fgfr1* expression in wild-type 6-somite-stage embryos. **d**, Lateral view (anterior, left) showing *fgfr1* expression in KV (bracket) and midbrain–hindbrain (red arrowhead). **e**, Tailbud showing *fgfr1* expression in KV (white arrow, dorsal view), presomitic mesoderm (red arrowhead) and lateral plate mesoderm (black arrow). **f**, **g**, *spaw* expression (arrows) in DFC$^{control\ MO}$ and DFC$^{fgfr1\ MO\text{-}1}$ at the 18–20-somite stage. **h**, Percentages of *spaw* expression in DFC and yolk MO-injected embryos. *spaw* was altered in DFC$^{fgfr1\ MO\text{-}1}$ ($n = 69$) versus DFC$^{control\ MO}$ ($P < 1.19 \times 10^{-5}$; $n = 121$), with no difference between yolk$^{control\ MO}$ ($n = 57$) and yolk$^{fgfr1\ MO\text{-}1}$ ($P < 0.90$; $n = 59$).

[1]Department of Neurobiology and Anatomy, University of Utah School of Medicine, Eccles Institute of Human Genetics, Building 533, Room 3160, 15 North 2030 East, Salt Lake City, Utah 84112-5330, USA. †Present address: Department of Cell and Developmental Biology, SUNY Upstate Medical University, 750 East Adams Street, Syracuse, New York 13210, USA.

of knockdown of Fgfr1 in yolk alone (yolk$^{fgfr1\ MO\text{-}1}$) were similar to those in yolk$^{control\ MO}$ (Fig. 1h; $P < 0.90$). These results indicate that cell-autonomous Fgfr1 signalling in DFC/KV cells is necessary for asymmetric expression of *spaw* in lateral plate mesoderm.

What role does Fgfr1 signalling have in DFC/KV function? Atypical protein kinase C (aPKC), an apical marker of polarized KV epithelial cells[16], revealed that KV were of normal size and shape in *fgfr1* morphants (Fig. 2a, b; $n = 15/15$, control $n = 16/16$), in contrast to dismorphic KV phenotypes seen in *ntl* or *spadetail* (*spt*, also known as *tbx16*) mutants and morphants[16]. Thus, morphogenesis of the KV epithelium is not dependent on Fgfr1 signalling. However, KV cilia were shorter in *fgfr1* MO-1 (see Methods) compared to control morphants and wild-type embryos (Fig. 2a–c; $P < 1.9 \times 10^{-8}$); the number of cilia was unaltered (Fig. 2; $P < 0.98$). Similar results were obtained from *fgfr1* MO-2 (data not shown). Importantly, *Xenopus fgfr1* mRNA[17] rescued cilia defects induced by *fgfr1* MO-1 (Fig. 2c; $P < 4.70 \times 10^{-5}$), demonstrating that cilia defects in *fgfr1* morphants are specific to Fgfr1 knockdown.

Additional approaches were used to assess the requirement of FGFR signalling for normal KV cilia length. Zebrafish embryos treated during the shield stage with a pharmacological inhibitor of FGFR activity, SU5402 (refs 18 and 19), had shorter cilia compared to dimethylsulphoxide (DMSO)-treated controls (Fig. 2d; $P < 3.26 \times 10^{-6}$). Treatment at subsequent stages altered left–right development but not cilia length (J.M.N. and H.J.Y., manuscript in preparation), indicating that FGF signalling has multiple stage-specific roles in left–right development. We analysed transgenic embryos carrying a heat-shock-inducible dominant-negative *fgfr1* (*hsp70:dn-fgfr1*) fused to enhanced green fluorescent protein (eGFP), which identifies transgenic embryos from their non-transgenic siblings[20]. When DN-Fgfr1 was activated at 60% epiboly, transgenic embryos had shorter cilia compared to heat-shocked non-transgenic siblings (Fig. 2e; $P < 6.94 \times 10^{-3}$) and non-heat-shocked siblings (Fig. 2e; $P < 6.99 \times 10^{-3}$), both of which had normal length cilia (Fig. 2e; $P < 0.61$). Brief hyperactivation of FGF signalling by inducible Fgfr (iFgfr)[21] avoided overexpression defects but did not increase cilia length (Supplementary Fig. 3).

Which ligands signal through Fgfr1 to control cilia length? Fgf8 binds several FGFRs[22] and Fgfr1 morphants phenocopy midbrain–hindbrain defects seen in zebrafish Fgf8 mutants (also known as *acerebellar*, *ace*)[5,23]. This indicates that Fgfr1 is a functional receptor for Fgf8 (ref. 23). *ace* mutants have left–right defects and a minority fail to form a KV lumen[5]. We found that *fgf8*-deficient embryos express KV differentiation markers (*sox17*, $n = 87/98$), form an epithelium with normal apical-basal polarity (aPKC, $n = 10/10$), and, despite 33% not filling the KV lumen, develop normal numbers of cilia with normal length (Fig. 2f; $P < 0.53$).

Another FGF ligand, *fgf24*, has overlapping expression with *fgf8* in and around DFC/KV cells[24]. *fgf24* mutants (*ikarus*, *ika*)[25] and siblings had normal length KV cilia (average cilia length = 6.2 μm; 498 cilia, 12 embryos). To test for redundant function of Fgf8 and Fgf24, we injected *fgf24* MO into *ace* mutants to reduce the amount of Fgf8/Fgf24 activity. *ace* heterozygotes injected with *fgf24* MO had shorter KV cilia than uninjected *ace* heterozygotes (Fig. 2f; $P < 0.015$), and *ace* homozygotes injected with *fgf24* MO had KV cilia lengths comparable to those of *fgfr1* morphants (Fig. 2f; $P < 3.63 \times 10^{-7}$). Similarly, *ika* mutants injected with *fgf8* MO had shorter cilia (Supplementary Fig. 4). Wild-type, *ika* mutants and siblings injected with *fgf24* MO had normal length cilia (Fig. 2f; $P < 0.28$), arguing against off-target MO effects. These results indicate that Fgf8 and Fgf24 ligands function, probably through Fgfr1, to control cilia length. Thus, results from MOs against Fgfr1, pharmacological inhibitors of FGFRs, transgenic expression of DN-Fgfr1, and mutants and MOs of multiple FGF ligands indicate that FGF signalling is necessary to control KV cilia length.

To assess whether cilia-driven directional fluid flow in KV was altered by the cilia defects in *fgfr1* morphants, we tracked movement of fluorescent beads injected into the lumen of KV[13]. In control morphants, fluorescent beads had a persistent counter-clockwise

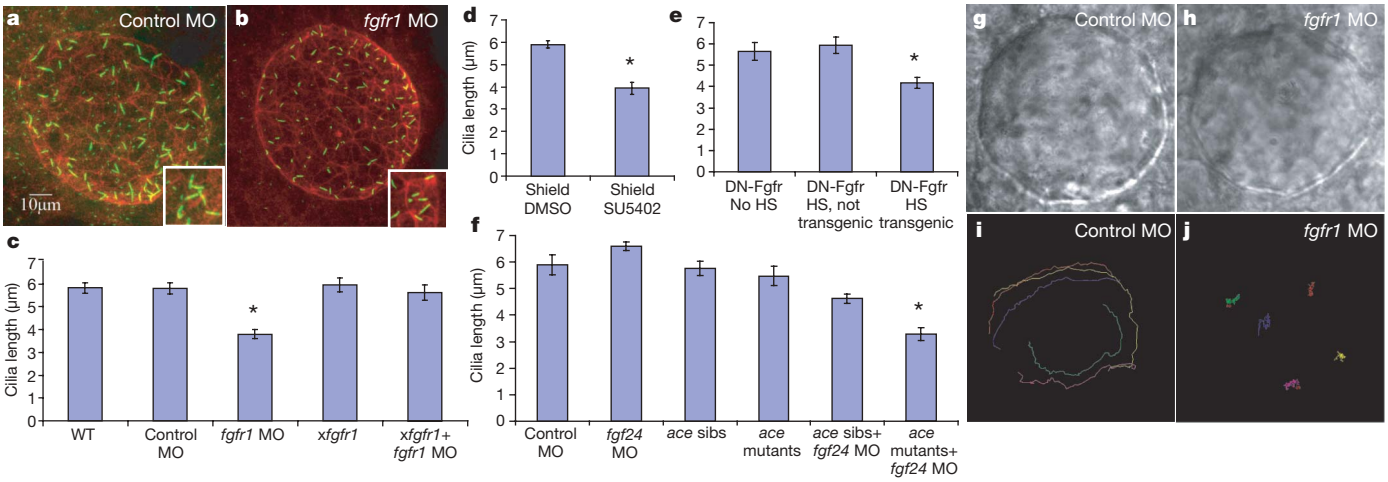


Figure 2 | FGF signalling controls cilia length and directional fluid flow in Kupffer's vesicle. **a**, **b**, Confocal images of 10-somite-stage embryos with the KV labelled with antibodies against aPKC (red) and acetylated tubulin (green). Control and *fgfr1* MOs had similar KV structure, but cilia were shorter in *fgfr1* MOs (compare insets in **a** and **b**). **c**, Cilia lengths were significantly different ($P < 2.88 \times 10^{-6}$) in *fgfr1* MOs (688 cilia; 18 embryos) versus control MOs (437 cilia; 9 embryos). Cilia length was similar in wild-type (WT) uninjected (533 cilia; 10 embryos) and control MO ($P < 0.93$), and cilia numbers per KV were similar in control and *fgfr1* MOs ($P < 0.26$). Cilia length defects in *fgfr1* MOs were rescued by *Xenopus fgfr1* (x*fgfr1*) mRNA ($P < 4.70 \times 10^{-5}$; 807 cilia; 21 embryos). Injection of x*fgfr1* mRNA alone had no effect on cilia length ($P < 0.73$; 526 cilia, 14 embryos). **d**, Embryos treated with SU5402 during shield stage (248 cilia; 12 embryos) had shorter cilia compared to DMSO control embryos ($P < 3.26 \times 10^{-6}$; 686 cilia; 15 embryos). **e**, Cilia were shorter in transgenic *hsp70:dn-fgfr1* embryos that were heat shocked (HS) at 60% epiboly (656 cilia; 19 embryos) compared to heat-shocked non-transgenic siblings ($P < 6.94 \times 10^{-3}$; 375 cilia; 10 embryos) and non-heat-shocked siblings ($P < 6.99 \times 10^{-3}$; 910 cilia; 16 embryos). **f**, There was no difference in cilia length ($P < 0.28$) in *fgf24* MOs (455 cilia; 10 embryos) versus control MOs (481 cilia; 10 embryos). However, cilia were shorter when both Fgf8 and Fgf24 ligands were diminished (*fgf24* MO in *ace* mutants; 12 embryos; 244 cilia), compared to single-ligand knockdown (*ace* mutants: $P < 1.39 \times 10^{-4}$; 10 embryos; 480 cilia; *fgf24* MO in *ace* siblings (sibs): $P < 3.44 \times 10^{-4}$; 15 embryos; 643 cilia) and wild-type *ace* siblings ($P < 3.63 \times 10^{-7}$; 13 embryos; 626 cilia). **g**, **h**, Differential interference contrast (DIC) images of KVs in control and *fgfr1* MOs injected with fluorescent beads. **i**, **j**, Bead paths tracked by Metamorph software. Directional KV fluid flow was absent in *fgfr1* MOs (**j**; $P < 6.4 \times 10^{-15}$; 44 beads, 9 embryos) compared to counter-clockwise flow in control MOs (**i**; 39 beads, 8 embryos). Error bars, s.e.m. Asterisks indicate conditions with statistically shorter cilia.

directional flow (Fig. 2i and Supplementary Movie 1). In contrast, beads in *fgfr1* morphants had no persistent directional flow (Fig. 2j and Supplementary Movie 2) indicating FGF signalling controls left–right patterning by regulating cilia length and KV fluid flow before initiation of asymmetric *spaw* expression.

The discovery that FGF signalling has a role in left–right patterning by regulating cilia indicates that other developmental roles attributed to FGF signalling might be due to cilia defects. To determine whether FGF-dependent regulation of cilia length is a more general developmental mechanism, we examined cilia in two epithelia that express Fgfr1, the pronephric ducts and ear (otic vesicle; Supplementary Fig. 5b, c). Pronephric ducts are primitive excretory organs containing motile cilia[14]. Inhibition of FGF signalling during *Xenopus* embryogenesis inhibits pronephric development[26], but no mechanism has been elucidated. Pronephric duct cilia at 26-somite stage were shorter in *fgfr1* morphants than in wild-type embryos (Fig. 3a, b, e; $P < 4.24 \times 10^{-4}$). Consistent with pronephric cilia defects, *fgfr1* morphants develop cystic kidneys (Supplementary Fig. 6). In the zebrafish ear, two types of cilia are required for otolith formation: tethering cilia and motile cilia. Tethering cilia attract seeding granules and, when reduced in number or length, granules are not organized correctly for otolith formation[19]. In zebrafish, knockdown of Fgf8 or Fgfr1 perturbs otic vesicle and otolith formation[23], and the otic vesicle cilia number is altered when FGF signalling is pharmacologically inhibited[18]. Here, *fgfr1* morphants had shorter tethering cilia and otolith defects (Fig. 3c, d, Supplementary Fig. 6d, e; $P < 1.1 \times 10^{-7}$), indicating that the otic vesicle and otolith defects seen in *fgfr1* MO-1 are due to defects in cilia length. Thus, FGF signalling controls cilia length and function in multiple tissues during zebrafish development.

To explore whether control of cilia length by FGF signalling is conserved in vertebrates, two types of epithelial cilia were examined in *Xenopus laevis*: monocilia on gastrocoel roof plate (GRP) implicated in left–right patterning[4], and mucociliary epithelial cilia that move fluid across the external epidermis[3]. Because dn-fgfr1 causes gastrulation defects when expressed ubiquitously during early embryogenesis, we co-injected *dn-fgfr1* and GFP mRNA into cell lineages that contribute to either the GRP or the mucociliary epithelium (Supplementary Fig. 1d–f). GRP cells co-expressing GFP and dn-fgfr1 had shorter cilia compared to neighbouring GRP cells in the same embryo ($P < 6.0 \times 10^{-3}$; Fig. 3i, j, m) and GRP cells in embryos expressing GFP alone ($P < 2.7 \times 10^{-3}$; Fig. 3g, h, m). In mucociliary epithelium, cells co-expressing GFP and dn-fgfr1 had shorter cilia than cells expressing GFP alone ($P < 0.019$; Fig. 3k, l, n). These results indicate that FGF signalling controls cilia length in diverse epithelia, and suggests that the regulation of cilia length by FGF signalling is evolutionarily conserved.

To address how Fgfr1 regulates cilia length, we analysed cell differentiation, epithelial cell polarization and cilia formation of KV cells in zebrafish[16]. In *fgfr1* morphants, two markers of the DFC/KV cell lineage, *sox17* (ref. 12) and *dnah9* (ref. 13), showed similar expression in wild type and *fgfr1* morphants, indicating correct DFC/KV cell differentiation (Fig. 4a–d, i). The apical membrane marker aPKC and tight junction marker ZO-1 revealed that apical–basal polarity in KV cells was intact in *fgfr1* morphants compared to wild-type controls (Supplementary Fig. 7a–d). Furthermore, cilia in *fgfr1* morphants were correctly positioned at the apical surface facing the KV lumen (Supplementary Fig. 7e, f). In contrast to the apparent normal differentiation and polarization of KV cells in zebrafish *fgfr1* morphants, two members of transcription factor families implicated in ciliogenesis[27,28], *foxj1* (also known as *hfh4*) and *rfx2* (B.W.B. and H.J.Y., manuscript in preparation), were downregulated (Fig. 4e, f, i). Correspondingly, expression of *ift88* (also known as *polaris*), an intraflagellar transport gene required for normal length cilia in zebrafish[29], was diminished in *fgfr1* morphants (Fig. 4g–i). Reduced *ift88* expression is consistent with intraflagellar-transport-defective phenotypes seen in *fgfr1* morphants, including curved body axis, kidney cysts and shortened cilia (Fig. 2a–f and Supplementary Fig. 6). From these results, we propose that Fgf8 and Fgf24 activate Fgfr1 cell-autonomously in KV cells to maintain a transcriptional network that allows normal expression of intraflagellar transport proteins required for normal length cilia (Fig. 4j).

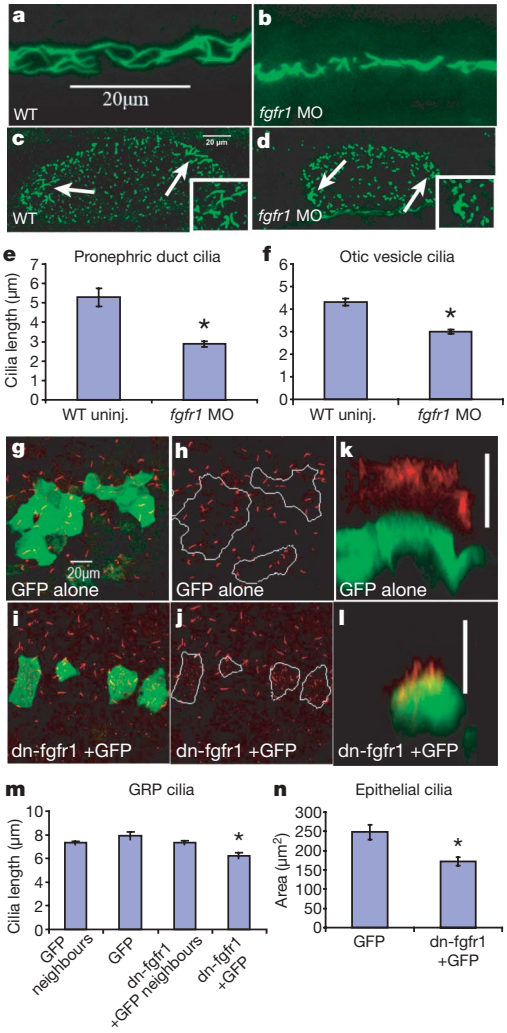


Figure 3 | Cilia length in pronephric ducts, otic vesicles, gastrocoel roof plate epithelia and mucociliary epithelia is controlled by FGF signalling. **a**, **b**, **e**, Pronephric duct cilia were shorter and disorganized in *fgfr1* MOs ($P < 4.24 \times 10^{-4}$; 528 cilia; 10 embryos) compared to wild-type (517 cilia; 10 embryos) 26-somite-stage embryos. **c**, **d**, **f**, Otic vesicle tethering cilia (arrows and inset) were shorter ($P < 1.10 \times 10^{-7}$) in *fgfr1* MOs (325 cilia; 10 embryos) compared to wild-type embryos (322 cilia; 8 embryos) at 24 hours post-fertilization (h.p.f.). **g–j**, **m**, GRP cilia in *Xenopus* embryos were normal length in cells expressing GFP alone (green cells in **g**, outlined in **h**; 316 cilia; 18 embryos, $P < 0.11$), neighbouring cells (outside boundaries in **h**; 653, 18 embryos) and cells neighbouring dn-fgfr1 + GFP expression (outside boundaries in **j**; 652 cilia, 15 embryos, $P < 0.99$). In contrast, GRP cilia were shorter in cells expressing dn-fgfr1 + GFP (**i**, inside boundaries in **j**; 155 cilia, 15 embryos) compared to neighbouring cells ($P < 6.1 \times 10^{-3}$) and cells expressing GFP alone ($P < 2.7 \times 10^{-3}$). **k**, **l**, *Z*-plane rendering of mucociliary epithelia (scale bar, 20 μm), showing shorter cilia in cells expressing dn-fgfr1 + GFP (13 cells, 7 embryos) compared to controls expressing GFP alone (14 cells, 4 embryos). **n**, Multicilia area is reduced in cells expressing dn-fgfr1 + GFP ($P < 0.019$). Error bars, s.e.m. Asterisks indicate conditions with statistically shorter cilia.

Monocilia are found on almost all cells and have been implicated as sites for receiving or modulating cell–cell signalling pathways such as hedgehog[1], platelet-derived growth factor (PDGF)[1] and Wnt[2]. Interactions among signalling pathways are of great interest in understanding how cells integrate diverse signals. Extrapolating from our discovery of a link between FGF signalling and cilia function in zebrafish and *Xenopus*, we propose that (1) some of the apparent interactions between FGF signalling and other cell signalling pathways might be due to FGF-dependent changes in cilia, which then influence the ability of

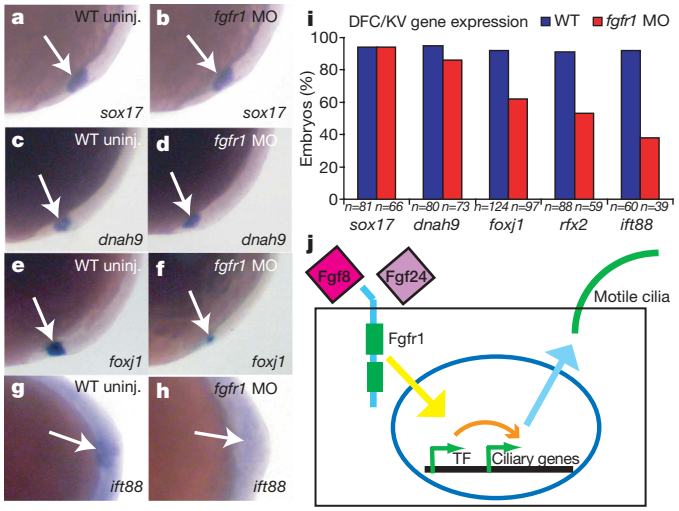


Figure 4 | FGF signalling controls ciliogenic genes in zebrafish DFC/KV cells. a, **b**, *sox17* expression in DFC/KV (and endoderm cells in a different focal plane) in 90% epiboly embryos was normal in *fgfr1* MOs and wild-type (WT) embryos. **c**, **d**, Expression of *dnah9* in 95% epiboly embryos was normal in *fgfr1* MOs and wild-type embryos. **e**, **f**, In contrast, *foxj1* was downregulated in *fgfr1* MOs versus wild-type embryos at 90% epiboly. **g**, **h**, Similarly, *ift88* was downregulated in *fgfr1* MOs versus wild-type embryos at tailbud stage. **i**, Comparison of percentage of embryos with wild-type expression levels of each gene indicated. **j**, Proposed mechanism by which FGF signalling controls the length of motile cilia: FGF ligands bind to Fgfr1, activating downstream transcription factors (TF) including *foxj1* and *rfx2*. These transcription factors activate intraflagellar transport genes (for example, *ift88*) to maintain motile cilia length on epithelial cells.

cells to receive and integrate other cell–cell signals, and (2) a spectrum of developmental defects and human diseases caused by defects in FGF signalling might be due to defects in cilia length or function.

METHODS SUMMARY

***Xenopus* mRNA injections.** For *Xenopus* GRP monocilia analysis, embryos were injected with 200 pg GFP mRNA alone (lineage tracer) or co-injected with 400 pg *dn-fgfr1* mRNA into two dorsal cells of a 32-cell embryo.

For *Xenopus* epithelial cell analysis, embryos were injected with 200 pg GFP mRNA alone or co-injected with 600 pg *dn-fgfr1* mRNA into a single ventral cell of a 16-cell embryo.

Statistics. Cilia measurements were analysed using a two-tailed Student's *t*-test, and analysis of *spaw* proportions were conducted using Fisher's exact test. In a given embryo, each cilium was measured in the tissue of interest and the average cilia length per embryo was determined. Averages for controls and experimentals were compared within each clutch of embryos. Outcomes were the same using a second analytical approach in which all cilia lengths were pooled and compared across all series of experiments. Analysis was done by R-Commander software package within the R Statistical Software platform[30]. Results are considered significant when $P < 0.05$ and results are expressed as mean ± s.e.m.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 July 2008; accepted 5 January 2009.
Published online 25 February 2009.

1. Eggenschwiler, J. T. & Anderson, K. V. Cilia and developmental signaling. *Annu. Rev. Cell Dev. Biol.* **23,** 345–373 (2007).
2. Gerdes, J. M. *et al.* Disruption of the basal body compromises proteasomal function and perturbs intracellular Wnt response. *Nature Genet.* **39,** 1350–1360 (2007).
3. Park, T. J., Mitchell, B. J., Abitua, P. B., Kintner, C. & Wallingford, J. B. Dishevelled controls apical docking and planar polarization of basal bodies in ciliated epithelial cells. *Nature Genet.* **40,** 871–879 (2008).
4. Schweickert, A. *et al.* Cilia-driven leftward flow determines laterality in *Xenopus*. *Curr. Biol.* **17,** 60–66 (2007).
5. Albertson, R. C. & Yelick, P. C. Roles for *fgf8* signaling in left–right patterning of the visceral organs and craniofacial skeleton. *Dev. Biol.* **283,** 310–321 (2005).
6. Boettger, T., Wittler, L. & Kessel, M. FGF8 functions in the specification of the right body side of the chick. *Curr. Biol.* **9,** 277–280 (1999).
7. Fischer, A., Viebahn, C. & Blum, M. FGF8 acts as a right determinant during establishment of the left–right axis in the rabbit. *Curr. Biol.* **12,** 1807–1816 (2002).
8. Meyers, E. N. & Martin, G. R. Differences in left–right axis pathways in mouse and chick: functions of FGF8 and SHH. *Science* **285,** 403–406 (1999).
9. Tanaka, Y., Okada, Y. & Hirokawa, N. FGF-induced vesicular release of Sonic hedgehog and retinoic acid in leftward nodal flow is critical for left–right determination. *Nature* **435,** 172–177 (2005).
10. Long, S., Ahmad, N. & Rebagliati, M. The zebrafish *nodal*-related gene *southpaw* is required for visceral and diencephalic left–right asymmetry. *Development* **130,** 2303–2316 (2003).
11. Roehl, H. & Nusslein-Volhard, C. Zebrafish *pea3* and *erm* are general targets of FGF8 signaling. *Curr. Biol.* **11,** 503–507 (2001).
12. Amack, J. D. & Yost, H. J. The T box transcription factor no tail in ciliated cells controls zebrafish left–right asymmetry. *Curr. Biol.* **14,** 685–690 (2004).
13. Essner, J. J., Amack, J. D., Nyholm, M. K., Harris, E. B. & Yost, H. J. Kupffer's vesicle is a ciliated organ of asymmetry in the zebrafish embryo that initiates left–right development of the brain, heart and gut. *Development* **132,** 1247–1260 (2005).
14. Kramer-Zucker, A. G. *et al.* Cilia-driven fluid flow in the zebrafish pronephros, brain and Kupffer's vesicle is required for normal organogenesis. *Development* **132,** 1907–1921 (2005).
15. Nonaka, S. *et al.* Randomization of left–right asymmetry due to loss of nodal cilia generating leftward flow of extraembryonic fluid in mice lacking KIF3B motor protein. *Cell* **95,** 829–837 (1998).
16. Amack, J. D., Wang, X. & Yost, H. J. Two T-box genes play independent and cooperative roles to regulate morphogenesis of ciliated Kupffer's vesicle in zebrafish. *Dev. Biol.* **310,** 196–210 (2007).
17. Amaya, E., Musci, T. J. & Kirschner, M. W. Expression of a dominant negative mutant of the FGF receptor disrupts mesoderm formation in *Xenopus* embryos. *Cell* **66,** 257–270 (1991).
18. Millimaki, B. B., Sweet, E. M., Dhason, M. S. & Riley, B. B. Zebrafish atoh1 genes: classic proneural activity in the inner ear and regulation by Fgf and Notch. *Development* **134,** 295–305 (2007).
19. Riley, B. B., Zhu, C., Janetopoulos, C. & Aufderheide, K. J. A critical period of ear development controlled by distinct populations of ciliated cells in the zebrafish. *Dev. Biol.* **191,** 191–201 (1997).
20. Lee, Y., Grill, S., Sanchez, A., Murphy-Ryan, M. & Poss, K. D. Fgf signaling instructs position-dependent growth rate during zebrafish fin regeneration. *Development* **132,** 5173–5183 (2005).
21. Pownall, M. E. *et al.* An inducible system for the study of FGF signalling in early amphibian development. *Dev. Biol.* **256,** 89–99 (2003).
22. Zhang, X. *et al.* Receptor specificity of the fibroblast growth factor family. The complete mammalian FGF family. *J. Biol. Chem.* **281,** 15694–15700 (2006).
23. Scholpp, S., Groth, C., Lohs, C., Lardelli, M. & Brand, M. Zebrafish *fgfr1* is a member of the *fgf8* synexpression group and is required for *fgf8* signalling at the midbrain–hindbrain boundary. *Dev. Genes Evol.* **214,** 285–295 (2004).
24. Draper, B. W., Stock, D. W. & Kimmel, C. B. Zebrafish *fgf24* functions with *fgf8* to promote posterior mesodermal development. *Development* **130,** 4639–4654 (2003).
25. Fischer, S., Draper, B. W. & Neumann, C. J. The zebrafish *fgf24* mutant identifies an additional level of Fgf signaling involved in vertebrate forelimb initiation. *Development* **130,** 3515–3524 (2003).
26. Urban, A. E. *et al.* FGF is essential for both condensation and mesenchymal–epithelial transition stages of pronephric kidney tubule development. *Dev. Biol.* **297,** 103–117 (2006).
27. Brody, S. L., Yan, X. H., Wuerffel, M. K., Song, S. K. & Shapiro, S. D. Ciliogenesis and left–right axis defects in forkhead factor HFH-4-null mice. *Am. J. Respir. Cell Mol. Biol.* **23,** 45–51 (2000).
28. Bonnafe, E. *et al.* The transcription factor RFX3 directs nodal cilium development and left–right asymmetry specification. *Mol. Cell. Biol.* **24,** 4417–4427 (2004).
29. Bisgrove, B. W., Snarr, B. S., Emrazian, A. & Yost, H. J. Polaris and Polycystin-2 in dorsal forerunner cells and Kupffer's vesicle are required for specification of the zebrafish left–right axis. *Dev. Biol.* **287,** 274–288 (2005).
30. The R Development Core Team. *The R Foundation for Statistical Computing* ⟨http://www.r-project.org/foundation/⟩ (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Moon and M. Condic for critical discussions on the manuscript; M. Karthikeyan, J. Shen, D. Coombs and E. Martini for technical help; and S. Miyagawa-Tomita, K. Poss and H. Issacs for reagents. This work was supported by American Heart Association predoctoral fellowship to J.M.N., NRSA Postdoctoral fellowship to J.D.A. and grants from NHLBI, NICHD and Primary Children's Medical Foundation to H.J.Y.

Author Contributions J.M.N. performed all zebrafish experiments except KV flow analysis (by J.D.A.) and *Xenopus* experiments (by A.G.P.). B.W.B. cloned zebrafish *foxj1*, *rfx2* and *ift88*. J.M.N. and H.J.Y. wrote the manuscript with input from all co-authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.J.Y. (jyost@genetics.utah.edu).

METHODS

Zebrafish and *Xenopus* embryo culture. Oregon AB wild-type zebrafish (*Danio rerio*) were collected from natural matings, and were injected, raised and staged as described previously[13]. Heterozygote crosses with *ace*^{ti282a}, *fgf24*^{t22030} and *hsp70:dn-fgfr1* were used to produce *ace* and *fgf24* homozygous mutant embryos and *hsp70:dn-fgfr1* transgenic embryos, respectively[5,20,23,25]. *hsp70:dn-fgfr1* embryos from heterozygote crosses were incubated at 28 °C (no heat-shock activation) or at 60% epiboly for one hour at 37 °C (heat-shock activation) and then returned to 28 °C until collected for immunohistochemistry (IHC). *Xenopus* embryos were obtained using standard methods as previously described[3].

Morpholino and mRNA injections. Antisense MOs were obtained from Gene Tools, LLC and Open Biosystems. Fluorescently labelled MOs against Fgfr1 were designed using previously described sequences: translation-blocking 3-carboxyfluorescein-labelled *fgfr1* MO-1 (5′-GCAGCAGCGTGGTCTTCAT-TATCAT-3′)[23,31], translation-blocking 3-carboxyfluorescein-labelled *fgfr1* MO-2 (5′-CAAAGATCCTCTACATCTGAACTCC-3′)[31]. The *fgf24* MO (5′-AGGAGACTCCCGTACCGTACTTGCC-3′) and the 3-lissamine-labelled *fgf8* MO (5′-TAGGATGCTCTTACCATGAACGTCG-3′) have also been described previously[24,32]. Fluorescein-labelled standard negative control (5′-CCTCT-TACCTCAGTTACAATTTATA-3′) from Gene Tools, LLC was used in control injections. MO was injected into 1–4-cell zebrafish embryos for whole-embryo protein knockdown experiments[13]. A volume of 1 nl was delivered containing 5 ng of *fgf24* MO, 4 ng of *fgf8* MO, 4 ng of *fgfr1* MO-1, 8 ng of *fgfr1* MO-2, or 4 ng of control MO. For DFC^{MO} experiments, fluorescent MO was injected into the yolk of embryos at the 500–1,000-cell stage and embryos were selected by fluorescent microscopy for MO accumulation in DFC as described previously[12]. To control for activity of the protein of interest in the yolk alone, we used yolk^{MO} control injections: fluorescent MO was injected into dome stage to 30% epiboly stage embryos, and embryos were selected by fluorescent microscopy for MO diffusion throughout the yolk. For DFC^{MO} and yolk^{MO} injections, 1 nl was delivered containing 2 ng of *fgfr1* MO-1 or 2 ng of control MO. Capped x*fgfr1*, *dn-fgfr1*, *ifgfr*[21] and GFP mRNAs were made from linearized plasmid using the using the mMessage machine SP6 transcription kit (Ambion)[17]. For MO rescue experiments, 100 pg of x*fgfr1* was injected alone or co-injected with 5 ng of *fgfr1* MO-1 into 1–4-cell-stage zebrafish embryos. For iFgfr experiments, 2.5 pg of *ifgfr* mRNA was injected into 1–4-cell-stage zebrafish embryos.

***In situ* hybridization.** Digoxigenin RNA probes were generated using a Roche DIG RNA labelling kit. Complementary DNA templates used include *spaw*[10], *shh*[13], *ntl*[12], *fgfr1* (ref. 23), *sox17* (ref. 12), *pea3* (ref. 11), *erm*[11], *lefty1* (ref. 13), *dnah9* (ref. 13), *ift88* (ref. 29), *foxj1* (B.W.B., unpublished) and *rfx2* (B.W.B., unpublished). *In situ* hybridizations were performed as described previously[13], with automated wash and antibody incubation using a Biolane HTI machine (Huller and Huttner HG). After post-fixation, embryos were cleared in 100% EtOH for imaging. Embryos were stored in 70% glycerol and images were obtained and processed using a Nikon Coolpix5000 camera and Photoshop software (Adobe).

Immunofluorescence microscopy. For zebrafish IHC, embryos were fixed in 4% paraformaldehyde at 4 °C, dehydrated in a MeOH series, stored in 100% MeOH, rehydrated, boiled in 1 mM EDTA for five minutes (except IHC for pronephric cilia), and subsequently blocked for 1 h in PBS containing 5% sheep serum, 1% BSA, 1% DMSO and 0.1% Triton-X. Embryos were incubated in primary antibody including mouse anti-acetylated tubulin (1:300, Sigma T-6793), rabbit anti-atypical protein kinase C ζ (1:100; Santa Cruz sc-216) and mouse anti-ZO-1 (1:150; Zymed 33-9100). After washes with PBS containing 0.1% Triton-X, 1% DMSO and 1% BSA, embryos were blocked for 1 h and incubated in secondary antibody, including goat anti-rabbit Alexa Fluor 647 and goat anti-mouse Alexa Fluor 488. Embryos were cleared and mounted in Slow Fade Reagent (Molecular Probes). Images were acquired using an Olympus Fluoview FV300 laser scanning confocal microscope and assembled using ImageJ (NIH) and Photoshop (Adobe) software. Confocal *Z*-series images were assembled to present the sum of the focal planes; cilia length was measured using Metamorph software (Universal Imaging Corp).

For GRP monocilia imaging, injected *Xenopus* embryos were collected at Nieuwkoop & Faber stage 17 (ref. 33), and the vitelline membrane removed, fixed overnight in 4% PFA in PBS, dehydrated in methanol and stored at −20 °C. Embryos were dissected following rehydration to expose GRP cilia according to previous methods[4]. For epithelial cilia analysis, injected embryos were collected at stage 26 and kept whole[3]. Embryos were blocked in 10% lamb serum in PBS/0.1% Triton-X (PBST), with PBST-only washes. Cilia were labelled as for zebrafish and injected cells were visualized using a polyclonal GFP antibody (1:400; Torrey Pines Biolabs). Anti-mouse Alexa Fluor 568 and anti-rabbit Alexa Fluor 488 secondary antibodies were used. Samples were mounted in PBST and imaged using an Olympus Fluoview FV300 confocal microscope. To measure epithelial cilia length, images were processed using Fluoview software to render the cilia in the *x*–*z* plane and then images and cilia length for both epithelial and GRP cilia were measured as for zebrafish.

KV flow analysis. Embryos were dechorionated at 6–8-somite stage and mounted in 1% low melt agarose. Fluorescent beads (0.5–2 μm; Polysciences, Inc.) were injected into KV and imaged on a Leica DMRA compound microscope using a ×40 Plan Apo objective with a Coolsnap HQ digital camera (Photometrics), Metamorph (Universal Imaging Corp) to track individual beads and calculate velocity, and Quicktime (Apple) to display movies.

Pharmacological treatments. Shield-stage embryos were incubated in 24-well tissue culture dishes (25–30 embryos per well) in either SU5402 (Calbiochem)[18,19] resuspended in DMSO or AP20187 (Ariad) resuspended in EtOH, and diluted into embryo water to a concentration of 20–25 μM for SU5402 (concentration dependant on drug lot) or 1.25 μM for AP20187. For a vehicle control, an equivalent volume of DMSO or EtOH was added to embryo water. At after 1 h, embryos were washed with embryo water and incubated in the 24-well dishes until fixed for IHC.

31. Thummel, R. *et al.* Inhibition of zebrafish fin regeneration using in vivo electroporation of morpholinos against *fgfr1* and *msxb. Dev. Dyn.* **235**, 336–346 (2006).
32. Draper, B. W., Morcos, P. A. & Kimmel, C. B. Inhibition of zebrafish *fgf8* pre-mRNA splicing with morpholino oligos: a quantifiable method for gene knockdown. *Genesis* **30**, 154–156 (2001).
33. Nieuwkoop, P. D. & Faber, J. *Normal Table of Xenopus Laevis (Daudin): A Systematical and Chronological Survey of the Development From the Fertilized Egg Till the End of Metamorphosis* (Garland, 1994).

LETTERS

Clustering of $InsP_3$ receptors by $InsP_3$ retunes their regulation by $InsP_3$ and Ca^{2+}

Taufiq-Ur-Rahman[1], Alexander Skupin[2], Martin Falcke[2,3] & Colin W. Taylor[1]

The versatility of Ca^{2+} signals derives from their spatio-temporal organization[1,2]. For Ca^{2+} signals initiated by inositol-1,4,5-trisphosphate ($InsP_3$), this requires local interactions between $InsP_3$ receptors ($InsP_3Rs$)[3,4] mediated by their rapid stimulation and slower inhibition[4] by cytosolic Ca^{2+}. This allows hierarchical recruitment of Ca^{2+} release events as the $InsP_3$ concentration increases[5]. Single $InsP_3Rs$ respond first, then clustered $InsP_3Rs$ open together giving a local 'Ca^{2+} puff', and as puffs become more frequent they ignite regenerative Ca^{2+} waves[1,5–9]. Using nuclear patch-clamp recording[10], here we demonstrate that $InsP_3Rs$ are initially randomly distributed with an estimated separation of ~1 μm. Low concentrations of $InsP_3$ cause $InsP_3Rs$ to aggregate rapidly and reversibly into small clusters of about four closely associated $InsP_3Rs$. At resting cytosolic [Ca^{2+}], clustered $InsP_3Rs$ open independently, but with lower open probability, shorter open time, and less $InsP_3$ sensitivity than lone $InsP_3Rs$. Increasing cytosolic [Ca^{2+}] reverses the inhibition caused by clustering, $InsP_3R$ gating becomes coupled, and the duration of multiple openings is prolonged. Clustering both exposes $InsP_3Rs$ to local Ca^{2+} rises and increases the effects of Ca^{2+}. Dynamic regulation of clustering by $InsP_3$ retunes $InsP_3R$ sensitivity to $InsP_3$ and Ca^{2+}, facilitating hierarchical recruitment of the elementary events that underlie all $InsP_3$-evoked Ca^{2+} signals[3,5].

$InsP_3$-activated currents recorded from patches excised from the outer nuclear envelope of DT40 cells[10] expressing rat $InsP_3R$ type 3 ($InsP_3R3$) are entirely due to $InsP_3R3$ (Fig. 1). With 10 μM $InsP_3$ in the pipette solution the single channel open probability (P_o) was 0.44 ± 0.05 (mean ± s.e.m.; $n = 6$) and the mean open time (τ_o) was 11.9 ± 1.6 ms. The distribution of closed times (τ_c) had two components (Fig. 1d). Recordings in the on-nucleus configuration confirmed these results (data not shown). The results are consistent with the gating scheme shown in Fig. 1d (see Supplementary Methods).

The number of channels within a patch (1.34 ± 0.13, $n = 109$) can be estimated reliably from the largest multiple of simultaneous openings to the unitary current level (Fig. 1e and Supplementary Methods). The distribution of $InsP_3Rs$ in a patch is random: it is not significantly different from a Poisson distribution (χ^2, $P > 0.05$; Fig. 1f and Supplementary Table 1). Others suggested that $InsP_3Rs$ are clustered in the nuclear envelope[11,12], but it seems likely that in making repeated recordings from the same nucleus they stimulated nuclei with $InsP_3$ before recording, and thereby caused $InsP_3R$ clustering (see later).

Channel activity (P_o; Fig. 2a–c), but not the number of active $InsP_3Rs$ (Fig. 2d), increased with $InsP_3$ concentration (effective concentration for half-maximum response (EC_{50}) = 1.38 ± 0.03 μM for patches with one $InsP_3R$). There was more than one $InsP_3R$ in 57% of active patches, and each opened to the same single-channel conductance (γ) (Figs 1e and 2a), but NP_o (the overall channel activity) was less than expected from the summed behaviour of lone $InsP_3Rs$ (Fig. 2e). For multi-$InsP_3R$ patches, the sensitivity to $InsP_3$ of NP_o was also significantly reduced ($EC_{50} = 2.47 \pm 0.25$ μM for patches with three $InsP_3Rs$; Fig. 2c and Supplementary Table 2). These observations prompted us to ask whether $InsP_3Rs$ behave independently in such multi-$InsP_3R$ patches or whether they interact, like some ryanodine receptors[13,14]. For each of the four states in patches with three $InsP_3Rs$ (closed and 1, 2 or 3 simultaneously open $InsP_3Rs$), the single channel open probability (P_o) predicted from the binomial distribution matched the observed P_o (Fig. 2f and Supplementary Methods). Similar results were obtained for patches with different numbers of $InsP_3R3$s and for type 1 $InsP_3Rs$ (Supplementary Figs 1 and 2). At resting cytosolic [Ca^{2+}], therefore, each $InsP_3R$ in a multi-$InsP_3R$ patch behaves identically and opens independently.

Our results present a conundrum. How can randomly distributed $InsP_3Rs$ that open independently behave with such uniformity, and yet so differently from lone $InsP_3Rs$, when a patch fortuitously contains several $InsP_3Rs$? Recordings from *Xenopus* nuclei also suggest that the heterogenous behaviour of lone $InsP_3Rs$ becomes more uniform when patches contain several $InsP_3Rs$ (ref. 15). We suggest that $InsP_3$ causes $InsP_3Rs$ to cluster[16], and that clustered $InsP_3Rs$ are less active. To test this hypothesis, nuclei were bathed in $InsP_3$ (10 μM, 2 min) before forming seals for patch-clamp recording. In these paired experiments, the mean number of $InsP_3Rs$ per patch was unaffected by $InsP_3$ pre-treatment (Supplementary Table 1), confirming that $InsP_3$ neither inactivated $InsP_3Rs$ nor affected the area of membrane trapped beneath the patch. However, the distributions of $InsP_3Rs$ were very different before and after $InsP_3$ treatment (Fig. 3a). In naive nuclei $InsP_3Rs$ were randomly distributed (Fig. 3b), but their distribution after $InsP_3$ pre-treatment differed significantly from the Poisson distribution ($P < 0.05$): many patches had no $InsP_3Rs$, single $InsP_3Rs$ were under-represented, and several patches had unusually large numbers of $InsP_3Rs$ (Fig. 3c). This clustering of $InsP_3Rs$ was fully reversed within 8–10 min of removing $InsP_3$ (Fig. 3a, d). P_o of lone $InsP_3Rs$ from naive nuclei (0.44 ± 0.05, $n = 6$) was indistinguishable from P_o of the only lone $InsP_3R$ caught within a patch after $InsP_3$ pre-treatment (0.41). P_o for each $InsP_3R$ within a cluster was also indistinguishable for recordings from naive (0.24 ± 0.01, $n = 18$) and $InsP_3$-pre-treated nuclei (0.25 ± 0.01, $n = 18$). Furthermore, there was no decrease in P_o during recordings that outlasted the $InsP_3$ pre-treatment (Supplementary Fig. 3). Thus clustering, rather than $InsP_3$ per se, decreases P_o.

The decrease in P_o as $InsP_3Rs$ cluster is identical whether clustering is evoked by the application of $InsP_3$ to an isolated patch (Fig. 2e, h) or to the entire nucleus (Fig. 3e). Both reduce P_o to ~54% that of lone $InsP_3Rs$. The latter condition better replicates the situation *in vivo*, confirming that results with isolated patches (Figs 1 and 2) faithfully

[1]Department of Pharmacology, Tennis Court Road, Cambridge CB2 1PD, UK. [2]Mathematical Cell Physiology, Max Delbrück Centre for Molecular Medicine, Robert Rössle Str. 10, 13092 Berlin, Germany. [3]Helmholtz Centre Berlin for Materials and Energy, Glienicker Str. 100, 14109 Berlin, Germany.

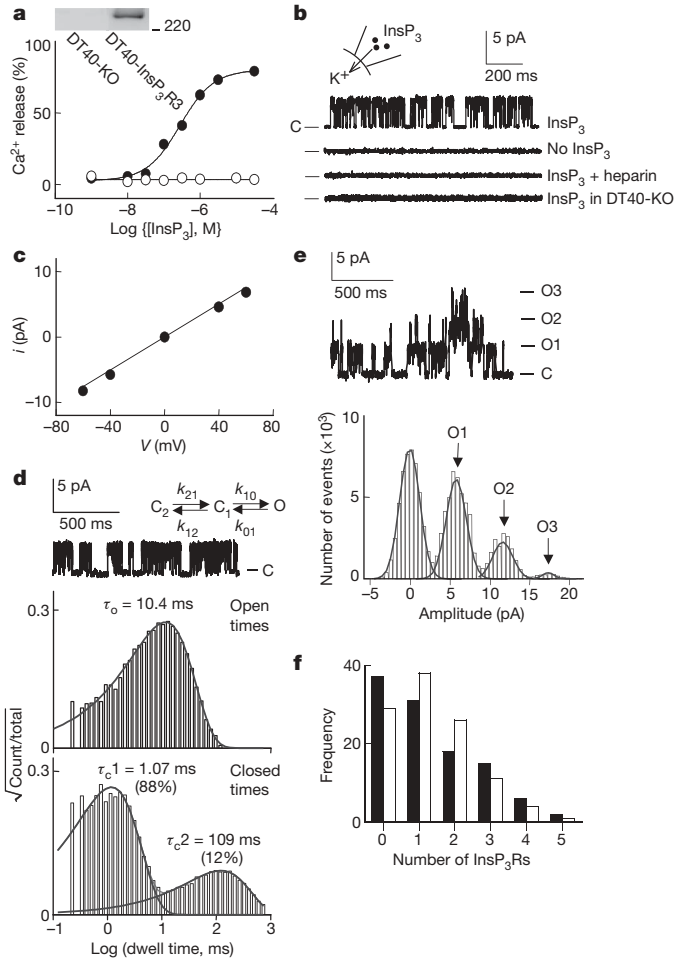


Figure 1 | InsP$_3$Rs are randomly distributed. **a**, InsP$_3$-evoked Ca^{2+} release from permeabilized DT40-InsP$_3$R3 (filled circles; $EC_{50} = 281 \pm 46$ nM, mean ± s.e.m.) and DT40-KO cells (open circles) ($n \geq 3$). Inset shows an immunoblot with InsP$_3$R3-specific antiserum (10 μg membrane protein per lane, a 220-kDa marker is shown). **b**, Currents recorded from excised patches with 10-μM InsP$_3$ in pipette solution. No currents were detected without InsP$_3$ ($n = 20$), with InsP$_3$ and heparin (100 μg ml^{-1}) ($n = 15$), or with InsP$_3$ in DT40-KO cells ($n > 30$). C denotes the closed state. **c**, The single-channel current–voltage (*i*–*V*) relationship for InsP$_3$-evoked current (K^+ conductance (γ_K) = 121 ± 2.8 pS, $n = 7$). **d**, Dwell-time distribution of a single InsP$_3$R3 stimulated with 10 μM InsP$_3$. Open time distribution of this typical recording is fitted with a single exponential function with $\tau_o = 10.4$ ms (mean = 11.9 ± 1.6 ms, $n = 6$). The probability density function for the τ_c distribution has two components ($\tau_c1 = 1.07$ ms, 88%, and $\tau_c2 = 109$ ms, 12%). Dwell-time distributions are consistent with the gating scheme (Supplementary Methods and Supplementary Figs 5 and 6). **e**, Typical all-points current amplitude histogram of an excised patch containing three InsP$_3$Rs stimulated with 10 μM InsP$_3$. C denotes the closed state. O1, O2 and O3 denote states with 1, 2 and 3 open channels. **f**, Observed (filled bars) and predicted (open bars) numbers of InsP$_3$Rs per patch from 109 patches (mean = 1.34) stimulated with 10–100 μM InsP$_3$.

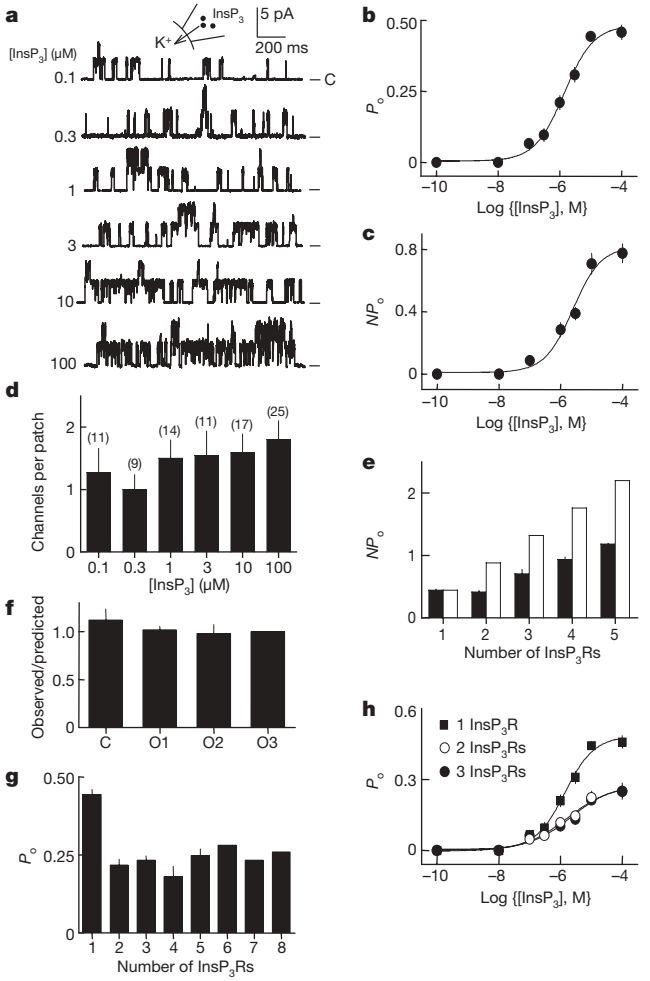


Figure 2 | Lone InsP$_3$Rs are more active than clustered InsP$_3$Rs at resting cytosolic [Ca^{2+}]. **a**, Typical records from patches (two InsP$_3$Rs per patch) stimulated with InsP$_3$. **b**, **c**, The effect of InsP$_3$ on P_o of patches containing a single InsP$_3$R (**b**) or on NP_o of patches with three InsP$_3$Rs (**c**) ($n \geq 4$). **d**, The numbers of InsP$_3$Rs detected in each patch for each InsP$_3$ concentration ($n = 9$–25). **e**, Predicted NP_o (NP_{lone}, open bars) and observed NP_o (filled bars) for patches containing 1–5 InsP$_3$Rs (n ≥ 3; n = 2 for the patch with 5 InsP$_3$Rs). **f**, For patches with three InsP$_3$Rs, the ratios of the observed to the predicted values are shown for the indicated numbers of simultaneous openings (Supplementary equation (4)). **g**, P_o as a function of the number of InsP$_3$Rs within a patch after stimulation with 10 μM InsP$_3$ (Supplementary equation (5)). **h**, The effect of InsP$_3$ on P_o for lone InsP$_3$Rs and for InsP$_3$Rs within multi-InsP$_3$R patches ($n \geq 4$). All error bars are s.e.m.

report the behaviour of InsP$_3$Rs roaming freely within the nuclear envelope. The effect of cluster size on P_o indicates that pairing of InsP$_3$Rs is sufficient to cause the maximal decrease in P_o. Additional InsP$_3$Rs can join a cluster, and their activity is attenuated, but InsP$_3$Rs within larger clusters are no more inhibited than pairs of InsP$_3$Rs (Fig. 2g, h and Supplementary Table 2). InsP$_3$Rs associate with actin[4] and microtubules[17], but neither is required for clustering-evoked changes in P_o (Supplementary Fig. 4).

To examine the effects of clustering on InsP$_3$R gating, we compared the mean open time (τ_o, Supplementary Information) of lone InsP$_3$Rs with τ_o for single channel openings from patches with several (*N*) InsP$_3$Rs (blue line in Fig. 3f). These τ_o should be similar if lone and grouped InsP$_3$Rs behave identically. For multi-InsP$_3$R patches, we also measured the duration of events in which all InsP$_3$Rs were simultaneously open ($\tau_{o,N}$, red line in Fig. 3f), and from that we calculated τ_o for individual, independently gated InsP$_3$Rs ($N\tau_{o,N}$). Both analyses gave the same result: τ_o for InsP$_3$Rs within a cluster was reduced to 47% of that for lone InsP$_3$Rs (Fig. 3f). A similar analysis of closed states confirmed that neither was affected by clustering (Supplementary Fig. 5 and Supplementary Table 3). InsP$_3$-evoked clustering almost doubles the rate of channel closure ($1/\tau_o$) and this alone is sufficient (Supplementary Fig. 6 and Supplementary Table 4) to account for the decreased P_o of clustered InsP$_3$Rs (Fig. 2g). Clustered InsP$_3$Rs open for half as long as lone InsP$_3$Rs (5.4 versus 11.9 ms), and pairing of InsP$_3$Rs is enough to cause the full effect (Fig. 2g). Other regulators of InsP$_3$Rs usually influence τ_c and so rates of channel opening[4]. The difference is important because τ_o will affect the time course of the initial Ca^{2+} release within elementary events[7] and thereby Ca^{2+}-mediated interactions between clustered InsP$_3$Rs. This is confirmed by simulations of intracellular Ca^{2+} spikes, in which the ~50% decrease in τ_o of clustered InsP$_3$Rs causes the

frequency of Ca^{2+} spiking to decrease by fourfold (Supplementary Fig. 7).

Within a patch, cluster size is limited to the number of $InsP_3Rs$ fortuitously caught beneath the patch pipette, but the clusters are larger for nuclei pre-treated with bath-applied $InsP_3$ (Fig. 3c). This demonstrates that a maximal concentration of $InsP_3$ causes >93% of $InsP_3Rs$ to cluster (85 out of 91 $InsP_3Rs$ from 88 nuclei pre-treated with $InsP_3$), and the average cluster contains 4.25 ± 0.38 $InsP_3Rs$ (Supplementary Methods). Inhibition of $InsP_3Rs$ within a cluster is not caused by feedback inhibition[4] from Ca^{2+} passing through neighbouring $InsP_3Rs$. Both bathing and pipette solutions have the same $[Ca^{2+}]$ and are buffered with BAPTA, the inhibition occurs at positive (Fig. 2) and negative holding potentials (Supplementary Discussion), and clustered $InsP_3Rs$ open independently (Fig. 2f). Because permeating ions cannot regulate neighbouring $InsP_3Rs$ under our recording conditions, inhibition must be mediated by contacts between $InsP_3Rs$. From this, we estimate that the average separation of $InsP_3Rs$ falls from ~1 μm to ~20 nm after clustering, and that clusters are ~2 μm apart (Supplementary Discussion). These spacings concur with confocal measurements suggesting that a Ca^{2+} puff originates from a cluster ~50 nm wide and that clusters are ~3 μm apart[18]. When expressed at high densities, $InsP_3Rs$ (ref. 19) and ryanodine receptors[20] form arrays with each tetrameric receptor contacting four others. We speculate that $InsP_3$-evoked clusters (of 4.25 ± 0.38 $InsP_3Rs$) exploit similar contacts and so, with single $InsP_3Rs$, form the fundamental units of Ca^{2+} signalling (Fig. 3g).

$InsP_3$-evoked clustering is complete within seconds of stimulation with a maximal concentration of $InsP_3$ (Supplementary Fig. 3). To resolve the time course, we used photolysis of caged $InsP_3$ in the pipette solution to increase rapidly the $InsP_3$ concentration bathing $InsP_3Rs$ trapped beneath the patch pipette. $InsP_3Rs$ were initially quiescent and then rapidly activated when $InsP_3$ was photoreleased (Fig. 3h). Irrespective of the number of $InsP_3Rs$ caught within a patch, τ_o was initially similar for all $InsP_3Rs$ (~10 ms). It then remained stable for many minutes for lone $InsP_3Rs$ (11.4 ± 0.5 ms), but fell within 2.5 s to 5.8 ± 0.3 ms for patches containing more than one $InsP_3R$ (Fig. 3i and Supplementary Fig. 8). Using τ_o to report $InsP_3R$ clustering suggests that clustering is complete within 2.5 s of $InsP_3$ addition. A similar analysis of P_o suggests a half-time for clustering of ~1.5–2 s (Fig. 3i). Our evidence that clustering does not require the cytoskeleton together with measurements of $InsP_3R3$ mobility[21,22] suggest that diffusion alone may be sufficient to allow $InsP_3R3$ clustering within a few seconds (Supplementary Discussion).

We can define the $InsP_3$ sensitivity of clustering by measuring the extent to which P_o of each $InsP_3R$ within a multi-$InsP_3R$ patch ($P_o = NP_o/N$, Supplementary Information) falls below P_o of an identically stimulated lone $InsP_3R$ (P_{lone}). This demonstrates that $InsP_3R$ clustering ($EC_{50} < 300$ nM) is about ten times more sensitive to $InsP_3$ than is channel opening ($EC_{50} = 2.02$ μM, Fig. 3e). Steady-state exposure to the low $InsP_3$ concentrations that evoke Ca^{2+} puffs[5,7] would, by assembling $InsP_3R$ clusters, allow both generation of puffs and loss of Ca^{2+} blips[23].

Clustering moves $InsP_3Rs$ (~1 μm apart) from being insulated from their neighbours by Ca^{2+}-buffering to domains (~20 nm apart) in which they will instantly experience high local $[Ca^{2+}]$ whenever a neighbour opens[24] (Supplementary Fig. 7). So far (Figs 1–3) we have prevented such interactions by using K^+ as a charge carrier and recording at a free $[Ca^{2+}]$ (200 nM) that mimics a resting cell. Subsequent experiments include 1 μM free $[Ca^{2+}]$ with $InsP_3$ in the pipette solution to simulate the $[Ca^{2+}]$ near open $InsP_3Rs$. For simplicity we use K^+ as a charge carrier. With 1 μM $[Ca^{2+}]$ in the pipette solution, $InsP_3R$ activity was increased: P_o for lone $InsP_3Rs$ almost doubled, as τ_c decreased (Fig. 4a)[4]. Neither the number of $InsP_3Rs$ per patch (1.12 ± 0.24) nor their random distribution (Fig. 4b) was affected by Ca^{2+}, but the interaction between $InsP_3Rs$ was altered. Whereas clustering reduced the overall activity of $InsP_3Rs$ (NP_o) at resting $[Ca^{2+}]$ (Fig. 2e), the inhibition was reversed by increased $[Ca^{2+}]$, such that the collective activity of a pair of $InsP_3Rs$ (NP_o) was the same as that predicted from the summed activity of two lone $InsP_3Rs$ (Fig. 4c). This did not result from disaggregation of clusters because at increased $[Ca^{2+}]$, $InsP_3Rs$ no longer opened independently. In patches with two $InsP_3Rs$ (open-channel noise prevented analysis of larger clusters), open probabilities did not fit the binomial distribution (Fig. 4e)—double open and closed events were over-represented (Supplementary Fig. 9).

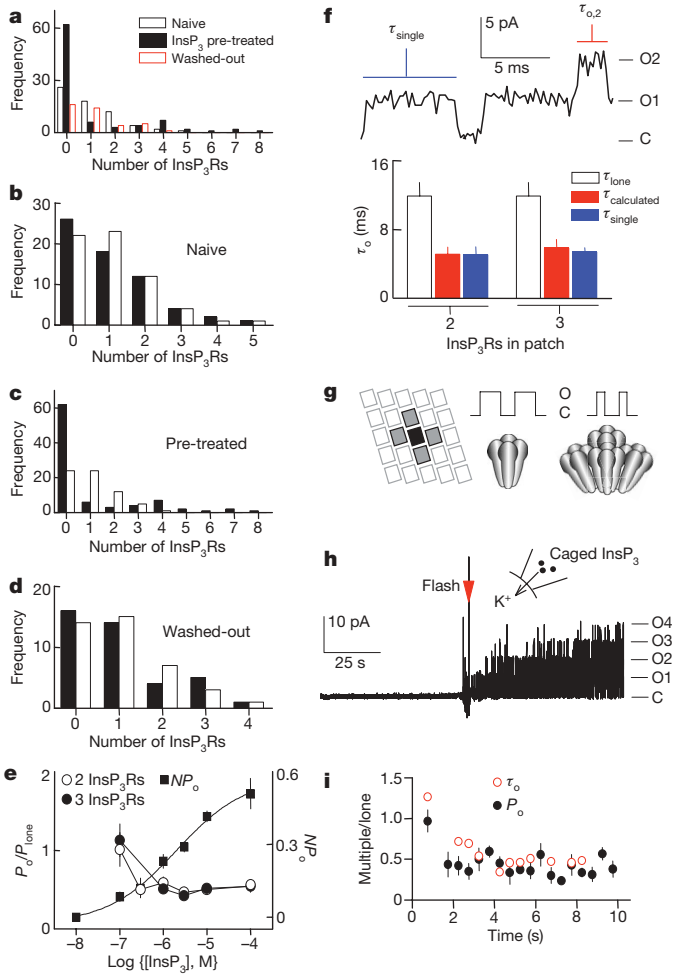


Figure 3 | Reversible clustering of $InsP_3Rs$ by $InsP_3$. **a**, The numbers of $InsP_3Rs$ detected in patches from naive nuclei ($n = 63$), after pre-treatment with bath-applied $InsP_3$ (10 μM, ~2 min; $n = 88$), or the latter after recovery for 8–10 min without $InsP_3$ (washed-out; $n = 40$). **b–d**, Observed (filled bars) and predicted (open bars) numbers of $InsP_3Rs$ per patch. **e**, The effects of $InsP_3$ on $InsP_3R$ clustering and gating. Clustering is reported by P_o/P_{lone} for patches with two or three $InsP_3Rs$, and gating by NP_o for patches with two $InsP_3Rs$ ($EC_{50} = 2.02 \pm 0.20$ μM). **f**, τ_o for patches with two or three $InsP_3Rs$ measured from the duration of single channel openings (blue line, τ_{single}) or calculated from the duration of openings to the Nth level (red line, $\tau_{calculated} = N\tau_{o,N}$). These are compared with τ_o for lone $InsP_3Rs$ (τ_{lone}). A typical trace is shown from a patch with two $InsP_3Rs$. **g**, $InsP_3$ drives $InsP_3Rs$ into small clusters consistent with the arrays (grey) formed by $InsP_3Rs$ at high density[19]. Within a cluster, each $InsP_3R$ opens independently, but closes more rapidly than a lone $InsP_3R$. **h**, A typical recording from a patch containing four $InsP_3Rs$ with $InsP_3$ released by flash photolysis from caged $InsP_3$ in pipette solution. Electrical noise caused by the flash is shown. **i**, From records similar to **h** (Supplementary Fig. 8), P_o (from NP_o/N) and τ_o were measured during each 0.5-s interval after the flash (1.5 s for the first interval). The ratio (multi-$InsP_3R$ patch/lone $InsP_3R$) is shown for both τ_o and P_o. Results (means ± s.e.m.) are from four (single) and seven (multiple, with 2–4 $InsP_3Rs$ per patch) patches.

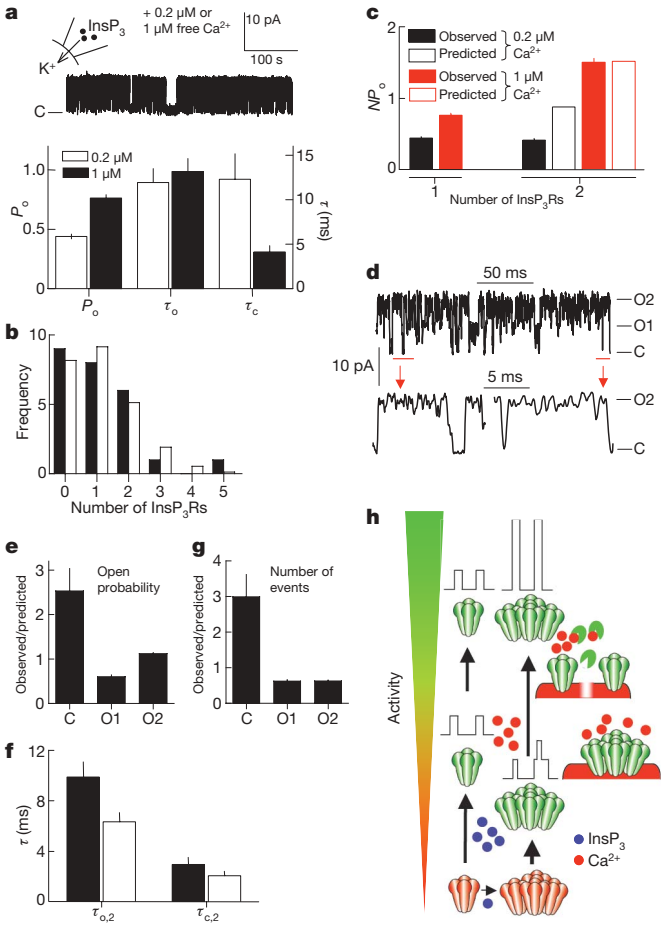


Figure 4 | Clustering retunes Ca^{2+} regulation of $InsP_3Rs$. **a**–**e**, Patches were stimulated with pipette solution containing 10 μM $InsP_3$ and (unless otherwise stated) 1 μM Ca^{2+}. **a**, A typical recording (top) and summary data (bottom; $n = 5$–6) from lone $InsP_3Rs$ show that increasing Ca^{2+} increases P_o by reducing τ_c. **b**, Observed (filled bars) and expected (open bars) numbers of $InsP_3Rs$ per patch. **c**, Observed and predicted NP_o for patches containing one or two $InsP_3Rs$ and stimulated with 10 μM $InsP_3$ in pipette solution containing 0.2 μM or 1 μM Ca^{2+} ($n = 5$–6). **d**, A typical recording from a patch with two $InsP_3Rs$, enlarged (red) to highlight transitions directly between closed (C) and double open (O2) states. **e**, The ratio of the observed to the predicted probability for closed (C) and single (O1) or double openings (O2) for patches with two $InsP_3Rs$ ($n = 6$; Supplementary equations (4) and (5)). **f**, Observed (filled bars) and expected (open bars) durations of events when both $InsP_3Rs$ are simultaneously open ($\tau_{o,2}$) or closed ($\tau_{c,2}$) for patches with two $InsP_3Rs$ ($n = 6$; Supplementary equations (6) and (7)). **g**, The ratio of the observed to the predicted numbers of transitions to each of the three states in a patch with two $InsP_3Rs$ ($n = 6$)[26]. **h**, At resting [Ca^{2+}], $InsP_3$ drives $InsP_3Rs$ into small clusters in which $InsP_3Rs$ gate independently, but with reduced P_o and $InsP_3$ sensitivity. Ca^{2+} reverses the inhibition imposed by clustering, openings within a cluster are more synchronized, and simultaneous openings are prolonged. Clustering primes $InsP_3Rs$ to respond by repressing their activity, and then allowing Ca^{2+} to unleash the coordinated gating of clustered $InsP_3Rs$ (Supplementary Fig. 7). All error bars are s.e.m.

Furthermore, there were many examples of $InsP_3Rs$ opening and closing directly to and from states with both $InsP_3Rs$ open (Fig. 4d). For paired $InsP_3Rs$, the double openings were prolonged by 50% (Fig. 4f), but were 47% less frequent than expected (Fig. 4g). The overall increase in P_o for double openings was therefore small (12%) and counteracted by a 39% decrease in the probability of only one $InsP_3R$ being open and a 116% increase in the probability of both being closed (Fig. 4e). Clustered $InsP_3Rs$ exposed to increased [Ca^{2+}] do not therefore behave independently. Their gating is coupled[13,14]: they are more likely to open and close together, and their simultaneous openings are prolonged (Supplementary Fig. 9). Coupled gating is not caused by local increases in cytosolic [Ca^{2+}], and must instead result from physical coupling of $InsP_3Rs$. However, under physiological conditions, clustered $InsP_3Rs$ are more likely to experience increased [Ca^{2+}] (because their neighbours may release it), and they are also tuned to respond most to it. By suppressing $InsP_3R$ activity at resting [Ca^{2+}], clustering increases the effect of a subsequent local increase in [Ca^{2+}] (Supplementary Fig. 7). Within a cluster, increased Ca^{2+} increases P_o (as it does for lone $InsP_3Rs$), but it also reverses the inhibition evoked by clustering and it causes coupled gating. These interactions exaggerate the effect of Ca^{2+} within a cluster (Fig. 4h). Thus, $InsP_3$ dynamically regulates both the assembly and behaviour of Ca^{2+} puff sites. $InsP_3$ rapidly drives $InsP_3Rs$ into small clusters, in which their $InsP_3$ and Ca^{2+} sensitivities are retuned to exaggerate Ca^{2+}-mediated recruitment of $InsP_3Rs$ and allow hierarchical recruitment of Ca^{2+} release events (Fig. 4h and Supplementary Fig. 7)[5,7].

METHODS SUMMARY

We established DT40 cell lines stably expressing rat $InsP_3R1$ or $InsP_3R3$. $InsP_3$-evoked Ca^{2+} release from the intracellular stores of permeabilized DT40 cells was measured using a low-affinity Ca^{2+} indicator (Mag-fluo-4) trapped within the endoplasmic reticulum. Nuclei were isolated by lysis of DT40 cells and allowed to adhere to a Petri dish coated with poly-D-ornithine. Patches excised from the outer nuclear envelope of these immobilized nuclei were used for patch-clamp recording[10]. K^+ was the charge carrier and, unless otherwise stated, all recordings were at +40 mV. For flash-photolysis experiments, the pipette solution contained 100 μM 'caged' $InsP_3$, from which $InsP_3$ was released by a single high-intensity flash from a Xe-flash lamp. Most analyses of currents used the QuB suite of programs (http://www.qub.buffalo.edu).

Received 23 August 2008; accepted 9 January 2009.
Published online 25 February 2009.

1. Berridge, M. J., Lipp, P. & Bootman, M. D. The versatility and universality of calcium signalling. *Nature Rev. Mol. Cell Biol.* **1**, 11–21 (2000).
2. Rizzuto, R. & Pozzan, T. Microdomains of intracellular Ca^{2+}: molecular determinants and functional consequences. *Physiol. Rev.* **86**, 369–408 (2006).
3. Marchant, J., Callamaras, N. & Parker, I. Initiation of IP_3-mediated Ca^{2+} waves in *Xenopus* oocytes. *EMBO J.* **18**, 5285–5299 (1999).
4. Foskett, J. K., White, C., Cheung, K. H. & Mak, D. O. Inositol trisphosphate receptor Ca^{2+} release channels. *Physiol. Rev.* **87**, 593–658 (2007).
5. Bootman, M. D., Berridge, M. J. & Lipp, P. Cooking with calcium: the recipes for composing global signals from elementary events. *Cell* **91**, 367–373 (1997).
6. Horne, J. H. & Meyer, T. Elementary calcium-release units induced by inositol trisphosphate. *Science* **276**, 1690–1694 (1997).
7. Marchant, J. S. & Parker, I. Role of elementary Ca^{2+} puffs in generating repetitive Ca^{2+} oscillations. *EMBO J.* **20**, 65–76 (2001).
8. Shuai, J., Rose, H. J. & Parker, I. The number and spatial distribution of IP_3 receptors underlying calcium puffs in *Xenopus* oocytes. *Biophys. J.* **91**, 4033–4044 (2006).
9. Sneyd, J. & Falcke, M. Models of the inositol trisphosphate receptor. *Prog. Biophys. Mol. Biol.* **89**, 207–245 (2005).
10. Dellis, O. *et al.* Ca^{2+} entry through plasma membrane IP_3 receptors. *Science* **313**, 229–233 (2006).
11. Mak, D.-O. D. & Foskett, J. K. Single-channel kinetics, inactivation, and spatial distribution of inositol trisphosphate (IP_3) receptors in *Xenopus* oocyte nucleus. *J. Gen. Physiol.* **109**, 571–587 (1997).
12. Ionescu, L. *et al.* Graded recruitment and inactivation of single $InsP_3$ receptor Ca^{2+}-release channels: implications for quantal Ca^{2+} release. *J. Physiol. (Lond.)* **573**, 645–662 (2006).
13. Marx, S. O. *et al.* Coupled gating between cardiac calcium release channels (ryanodine receptors). *Circ. Res.* **88**, 1151–1158 (2001).
14. Marx, S. O., Ondrias, K. & Marks, A. R. Coupled gating between individual skeletal muscle Ca^{2+} release channels (ryanodine receptors). *Science* **281**, 818–821 (1998).
15. Mak, D.-O. D. & Foskett, J. K. Effects of divalent cations on single-channel conduction properties of *Xenopus* IP_3 receptor. *Am. J. Physiol.* **275**, C179–C188 (1998).
16. Tateishi, Y. *et al.* Cluster formation of inositol 1,4,5-trisphosphate receptor requires its transition to open state. *J. Biol. Chem.* **280**, 6816–6822 (2005).
17. Bourguignon, L. Y., Iida, N. & Jin, H. The involvement of the cytoskeleton in regulating IP_3 receptor-mediated internal Ca^{2+} release in human blood platelets. *Cell Biol. Int.* **17**, 751–758 (1993).
18. Dargan, S. L. & Parker, I. Buffer kinetics shape the spatiotemporal patterns of IP_3-evoked Ca^{2+} signals. *J. Physiol. (Lond.)* **553**, 775–788 (2003).
19. Katayama, E. *et al.* Native structure and arrangement of inositol-1,4,5-trisphosphate receptor molecules in bovine cerebellar Purkinje cells as studied by quick-freeze deep-etch electron microscopy. *EMBO J.* **15**, 4844–4851 (1996).
20. Yin, C. C., Blayney, L. M. & Lai, F. A. Physical coupling between ryanodine receptor-calcium release channels. *J. Mol. Biol.* **349**, 538–546 (2005).

21. Fukatsu, K. *et al.* Lateral diffusion of inositol 1,4,5-trisphosphate receptor type 1 is regulated by actin filaments and 4.1N in neuronal dendrites. *J. Biol. Chem.* **279,** 48976–48982 (2004).
22. Ferreri-Jacobia, M., Mak, D.-O. D. & Foskett, J. K. Translational mobility of the type 3 inositol 1,4,5-trisphosphate receptor Ca^{2+} release channel in endoplasmic reticulum membrane. *J. Biol. Chem.* **280,** 3824–3831 (2005).
23. Sun, X.-P., Callamaras, N., Marchant, J. S. & Parker, I. A continuum of InsP_3-mediated elementary Ca^{2+} signalling events in *Xenopus* oocytes. *J. Physiol. (Lond.)* **509,** 67–80 (1998).
24. Falcke, M. Reading the patterns in living cells—the physics of Ca^{2+} signaling. *Adv. Phys.* **53,** 255–440 (2004).
25. Boehning, D., Joseph, S. K., Mak, D.-O. D. & Foskett, J. K. Single-channel recordings of recombinant inositol trisphosphate receptors in mammalian nuclear envelope. *Biophys. J.* **81,** 117–124 (2001).
26. Prole, D. L., Lima, P. A. & Marrion, N. V. Mechanisms underlying modulation of neuronal KCNQ2/KCNQ3 potassium channels by extracellular protons. *J. Gen. Physiol.* **122,** 775–793 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by The Wellcome Trust (C.W.T.), The Biotechnology and Biological Sciences Research Council (C.W.T.), a scholarship from the Jameel Family Trust (T.-U.-R.), and the IRTG 'Genomics and Systems Biology of Molecular Networks' of the Deutsche Forschungsgemeinschaft (A.S.). We thank S. Dedos for help with DT40 cells, D. Prole and B. Billups for advice, and T. Kurosaki for providing DT40-KO cells.

Author Contributions T.-U.-R. performed all experiments and, with C.W.T., analysed the data. A.S. and M.F. performed the modelling and contributed to discussions of diffusion. C.W.T. and T.-U.-R. wrote the paper with input from A.S. and M.F. The project was directed by C.W.T.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.W.T. (cwt1000@cam.ac.uk).

RETRACTION

doi:10.1038/nature07964

Remission in models of type 1 diabetes by gene therapy using a single-chain insulin analogue

Hyun Chul Lee, Su-Jin Kim, Kyung-Sup Kim, Hang-Cheol Shin & Ji-Won Yoon

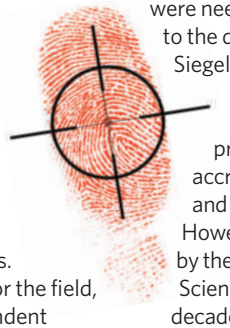
Nature 408, 483–488 (2000)

Three of the authors (H.C.L., K.-S.K. and H.-C.S.) wish to retract this Letter on the grounds that they have been unable to reproduce the results. The retraction has not been signed by Ji-Won Yoon (deceased) or by Su-Jin Kim, who maintains that the results are still valid.

Forensic evidence

Fresh career opportunities could develop in forensic science, if recommendations in a report from the US National Research Council are adopted, says forensic scientist and co-author Jay Siegel.

Forensic scientists need to prove their competence with recognized qualifications at different levels, says *Strengthening Forensic Science in the United States: A Path Forward*. Concerned members of Congress had asked the National Academy of Sciences to propose reforms that would coordinate and improve forensic-science analyses across federal, state and local jurisdictions. The report recommends mandatory certification for the pathologists, biologists, physicists, chemists and medical officers working in forensics.



CORBIS

To set these rigorous standards for the field, it calls for the creation of an independent National Institute of Forensic Science. Without such an institute, says report co-chair Constantine Gatsonis, a biostatistician at Brown University in Providence, Rhode Island, forensic science will continute to lack the funds needed to mature the field.

More thorough scientific evaluation of forensic protocols may generate new jobs, predicts Siegel, director of the forensic and investigative sciences programme at Purdue University in Indianapolis, Indiana. "The biggest problem in forensic science is a lack of science-based research to settle what can be considered evidence in the courtroom." For example, he says, despite the routine acceptance of fingerprints in the courts, evidence is still lacking as to how well a given fingerprint identifies a specific person.

Siegel believes that if Congress adopts some of the recommendations, the field will experience a hiring boom when the economy recovers. "There is a tremendous pent-up need for new scientists," says Siegel. A 2005 survey, *Census of Publicly Funded Forensic Crime Laboratories, 2002*, of crime-lab directors indicated 1,900 additional forensic scientists were needed to get case management down to the desired 30-day turnaround. And, Siegel says, staffing needs have only increased since then.

At present, certification programmes for individuals and accreditation of education programmes and crime laboratories are voluntary. However, these are not all supervised by the American Academy of Forensic Sciences (AAFS), which has spent the past decade establishing a board to examine the certifying bodies in current existence. Although AAFS president Thomas Bohan agrees that certification is important, he thinks the academy's existing system is sufficient. He believes that the report's emphasis on certification will prod most forensic scientists and institutions to flock to AAFS-approved certifying boards, making a new overseeing body unnecessarily complicated.

The recommendations could also push more forensic-science educational programmes to seek accreditation. Of the roughly 200 now operating, according to AAFS, only 19 are accredited by its Forensic Science Education Programs Accreditation Commission. ■

Virginia Gewin

POSTDOC JOURNAL

Job juggling

My wedding celebration is over, the flights booked, my visa nestled in my passport. Now all I have to do is complete all the projects I'm working on before I leave Australia to spend the next two years in the United States.

Since my postdoc contract ended last year, I have been paying the bills by working on three part-time projects that add up to a full-time workload. During an average day I juggle my time between them — from examining the impacts of dingos on Australia's mammals, to writing a book chapter on the impacts of climate change on Western Australian biodiversity, to writing website content for a new national climate-change research network.

I am grateful for the work, and dependent on the money it brings, but I yearn to do my own research. As I struggle to find enough hours in the day, unfinished manuscripts sit forlornly in a folder on my desktop. Others wait for me to address reviewer comments and resubmit them to journals. This does not bode well for my 2009 publication record.

With our forthcoming move to the United States, my husband working full time, and a toddler to care for, I can't see this cycle of part-time work ending any time soon. So perhaps I should embrace it rather than fight it.

Indeed, the benefits are many. I get the opportunity to work on a diverse range of interesting projects, and the flexible hours allow me more time with my son. And maybe one day I'll embrace those lonely manuscripts and finish them once and for all. ■

Joanne Isaac was a postdoc in climate-change effects on biodiversity at James Cook University, Townsville, Australia.

IN BRIEF

Drug firms cut back

Several drug research companies across North America, including five US-based firms and a Canadian biotech, have announced lay-offs.

Hospira of Lake Forest, Illinois, a pharmaceutical and medication delivery firm specializing in injectable drugs, will cut about 1,400 employees, or 10% of its global workforce. Cortex Pharmaceuticals of Irvine, California, which makes drugs to treat psychiatric and nervous-system disorders, cut 14 of 27 employees.

Poniard Pharmaceuticals of South San Francisco, California, is cutting eight of its 67 employees, discontinuing in-house preclinical research and focusing on picoplatin, a next-generation platinum chemotherapy. Adventrx Pharmaceuticals of San Diego, California, is cutting its payroll to five and is discontinuing drug-development efforts and business operations to focus on "strategic options". In December 2008 the company employed about 35 people, according to its website.

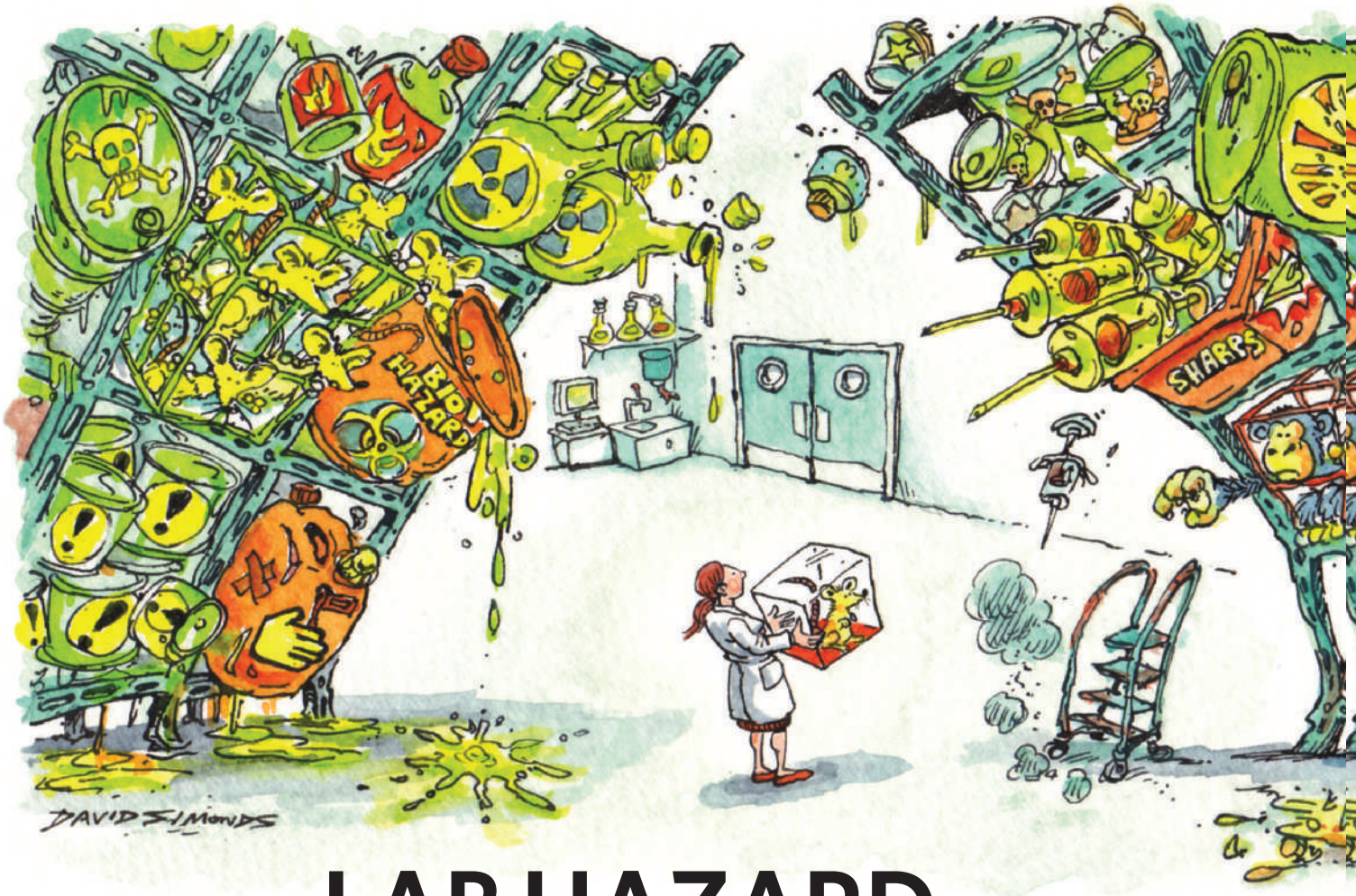
Synta Pharmaceuticals of Lexington, Massachusetts, cut 90 positions from its 220-member workforce owing to unfavourable late-stage clinical-trial results on a metastatic melanoma treatment. Synta has five programmes in clinical or preclinical development and several others in the discovery stage. Canada's Bellus Health is cutting its staff by nearly half. It did not report exact numbers, but the company employed 170 in December 2007, according to its website.

Syngene centre opens

Syngene International, a subsidiary of Indian biotech Biocon, and US drugmaker Bristol-Myers Squibb (BMS) have opened a research-and-development centre in Bangalore.

The 18,000-square-metre facility, which employs 270 researchers, helps advance BMS's discovery and early drug-development efforts. It will house 360 researchers by the end of the year and plans to ramp that number up to 450.

Work at the facility will span the drug discovery and development process. Construction began in March 2007, when BMS and Biocon agreed to focus on integrated drug discovery and development capabilities at Syngene.



D. SIMONDS

LAB HAZARD

Getting great results from experiments can be difficult, especially if the materials you work with decide to fight back. **Amber Dance** investigates some of the unappreciated risks of being at the bench.

No one told Karen Quigley her PhD project could make her ill. But during a clean-up of the rat room at Ohio State University in Columbus, Quigley had an asthma attack. She had developed an allergy to her rodent subjects — a health issue that persisted throughout her graduate studies and later postdoctoral work at Columbia University, New York. Quigley felt embarrassed wearing her "Darth Vader" get-up, a face mask with respirator, to handle the animals. "You need to do your work, so you just soldier on," says Quigley, who now sticks to humans at the US Department of Veterans Affairs New Jersey Healthcare System in East Orange. She enjoys her work, but would have liked to continue with animal models. Her allergy made that impossible.

Scientists are familiar with the clear dangers of laboratory work, such as concentrated acids and bases, sharp needles and radiation. But other health concerns are more subtle. Allergies and chemical sensitivities are a perennial threat to scientists who spend most of their time in the lab. Epidemiological data are scarce, so long-term hazards may be unknown. Even with known risks, those in a rush to collect data may skip safety measures. Generally they don't pay for their carelessness, but occasionally, lab health hazards can delay research, force a career move or cause serious injury or death (see 'Worst-case scenarios').

Sensitive subject

One study of animal-lab workers in Japan found that nearly a quarter of more than 5,000 survey respondents reported allergy symptoms (K. Aoyama *et al. Br. J. Ind. Med.* **49,** 41–47; 1992). Rodents are widespread culprits, but other mammals, insects and plants can also cause problems.

Allergic symptoms can curtail scientists' time at the bench; with continued exposure, asthma may also occur away from the lab. If an allergy is likely to develop, it usually arises within a few years of starting animal work. "It's fairly common, especially if people have an allergy already," says Christian Newcomer, executive director of the Association for Assessment and Accreditation of Laboratory Animal Care International in Frederick, Maryland. For some, avoiding the trigger may be the only option. "People do walk away from animal research because of this," Newcomer says.

Pharmacologist Mary Lynn Baniecki of Raleigh, North Carolina, tried everything to get around her rat allergy that developed during a master's project at Northeastern University in Boston, Massachusetts. She wore a mask and full-length disposable lab gown, and showered immediately after leaving the lab. But nothing helped. Her allergist told her to stop animal work to avoid becoming "severely ill", she says. The diagnosis forced her to abandon her research tracking dopamine neurons.

Animals are not the only potential health issue in a laboratory. Chemicals and other materials can cause allergies or sensitivity. For example, 8–12% of health-care workers are allergic to latex, according to the US National Institute for Occupational Safety and Health (NIOSH).

Baniecki discovered the hazards of chemicals at her next position. Leaving the rats behind, she embraced structure-based

pharmacology during a PhD at the State University of New York at Stony Brook. But two years into that work, she noticed the familiar signs — a tightness in her chest, a general sick feeling. By keeping a diary of activities and symptoms, she determined that she felt ill every time a colleague used gel electrophoresis to separate proteins. The source, Baniecki says, was the reducing agents used to break apart proteins. For those with a chemical sensitivity, the only answer is to minimize contact with the cause — not easy for something that is omnipresent in biochemistry labs. Baniecki had to rely on her colleagues to keep their reducing agents covered.

Chemical concerns

The working patterns of scientists make it difficult to collect epidemiological information on long-term health hazards. "It's the hazards that are not immediate that are the biggest problem," says Joe Crea, chief adviser on occupational hygiene for Safework South Australia in Adelaide. Scientists are a highly mobile population and health problems can go unreported. Diseases such as cancer can develop slowly so that by the time a person falls ill, there is no way to make a direct connection to their lab work. That makes it harder for safety managers to know exactly how dangerous certain chemicals may be.

Less than 2% of commercially available chemicals have been evaluated for carcinogenicity, according to the NIOSH. Instead, safety officers and scientists must make their best guess. It is assumed that ethidium bromide, a common reagent used to visualize DNA under ultraviolet light, is a carcinogen because it squeezes between the nucleotide bases of the genetic molecule, potentially causing mutations. Therefore, most researchers are careful to wear gloves when handling DNA gels, and some labs have switched to alternative DNA stains. Yet despite its suspected genome-altering properties, the scant data on ethidium bromide means it is not on the NIOSH list of potential occupational carcinogens.

Rapidly changing lab techniques, such as new synthesis procedures, can exacerbate safety concerns as chemicals with unknown long-term hazards come into vogue. Tim Brunker, an organic chemist at Towson University in Maryland, had no worries when synthesizing a new chemical entity during his postdoc at Dartmouth College in Hanover, New Hampshire. Brunker wore his usual lab attire — gloves and safety glasses with jeans and a T-shirt. After the preparation his nose itched, so he rinsed it. But after repeating the synthesis a few times, he broke out in a red, blotchy rash. It took a month of taking antihistamines around the clock for Brunker to recover. He stayed out of the lab, doing office work for much of that month, and avoided that chemical thereafter.

"In hindsight, I probably should have been more careful," Brunker says. "I never really thought that sort of thing could happen." These subtle health hazards are rarely discussed among lab workers. And researchers can view safety regulations as an exercise in overkill, particularly when even the lab supply of sodium chloride — table salt — comes with a warning that ingesting too much could be dangerous.

"The laboratory is ultimately a very safe place."
— Steve Benedict

The laboratory is ultimately a very safe place, says Steve Benedict, director of environment, health and safety at the University of California, San Diego. In 18 years of managing lab safety at San Diego and elsewhere, he has not encountered a serious incident or fatality.

Despite the potential hazards of chemicals and procedures, most accidents that befall researchers are not specific to lab work. Among scientists, there were three times as many workplace injuries from trips and falls than from harmful substances or environments in 2007, according to the US Bureau of Labor Statistics.

Benedict says his office can usually help scientists with a specific allergy or sensitivity to continue with their research. But sometimes the problem is too severe. Benedict's advice for those rare few: "Maybe the lab really isn't the place for you to be." ■

Amber Dance is a freelance writer based in Los Angeles, California.

WORST-CASE SCENARIOS

Laboratories are generally safe places, but sometimes a minor mistake can have consequences that go far beyond allergies or rashes. This is a selection of tragic incidents reported in the media over the past 15 years; they demonstrate the importance of taking great care in the lab.

A research assistant died in January this year from burns sustained in a university chemistry laboratory in California. Sheharbano Sangji had been working in the lab for only a few months when the plunger popped out of the syringe she was using to transfer tert-butyl lithium — which ignites spontaneously in air — causing her gloves and jumper to catch fire.

A chance splash of primate fluids cost research assistant Elizabeth Griffin her life in a 1997 incident at Yerkes Regional Primate Research Center research centre in Georgia. Because the rhesus macaques under study were caged, Griffin did not use safety glasses for the procedure being done. A piece of material contaminated with herpes B virus — probably urine or faeces — got into her eye and she died six weeks later.

Chemist Karen Wetterhahn spilt a drop of dimethylmercury on her gloved hand in 1996 at Dartmouth College in New Hampshire. At the time, it was not known that the chemical passes through latex, so she did not realize it had reached her skin. She fell ill five months later and died within the year.

In 2004, while drawing blood from Ebola-infected guinea pigs, Antonina Presnyakova accidentally stuck herself with the needle she was using. Presnyakova, a scientist at a virology laboratory in Russia, died two weeks later.

In 1994, a laboratory technician, working alone in a private lab in Perth, Australia, spilled hydrofluoric acid in his lap. He washed his limbs but did not apply calcium gluconate gel, the recommended treatment. The technician died 15 days later from multiple organ failure. **A.D.**

Caveat time traveller

Future-proof.

Gregory Benford

He was easy to spot — clothes from the twenty-first century, dazed look. I didn't have to say anything. He blurted out, "Look, I'm from the past, a time traveller. But I get snapped back there in a few minutes."

"I know." We stood in a small street at the edge of the city, dusk creeping in. Distant, glazed towers gleamed in the sunset and pearly lights popped on down along the main road. Jaunters always chose to appear at dawn or dusk, where they might not be noticed but could see a town. No point in transporting into a field somewhere, which could be any time at all, even the far past. Good thing he couldn't see the city rubble, too. Or realize this was how I made my living.

His mouth twisted in surprise. "You do? I thought I might be the first to come here. To this time."

I gave him a raised eyebrow. "No. There was another last week."

"Really? The professor said the other experiments failed. They couldn't prove they'd been into the future at all."

They always want to talk, though they'd learn more with their mouths closed.

He rattled on, "I have to take something back, to show I was here. Something —"

"How about this?" I pulled out a slim metal cylinder. "Apply it to your neck five times a day and it extracts cancer precursors. In your era, that will extend your average lifetime by several years."

His eyebrows shot up. "Wow! Sure —" He reached for it but I snatched it back.

"What do I get in exchange?" I said mildly.

That startled him. "What? I don't have anything you could use …" He searched his pockets in the old fashioned wide-label jacket. "How about money?" A fistful of bills.

"I'm not a collector, and those are worthless now, inflated away in value."

The time jaunter blinked. "Look, this is one of the first attempts to jump forward and back. I don't have —"

"I know, we've seen jaunters from your era already. Enough to set up a barter system. That's why I had this cancer-canceller."

Confusion swarmed in his face. "Lady, I'm just a guinea pig here. A volunteer. They didn't give me —"

I pointed. "Your watch is a pleasant anachronism, I'll take that." I gave him the usual ceramic smile.

He sighed with relief. "Great —" But I kept the cylinder away from him.

JACEY

"That's an opener offer, not the whole deal." A broader smile.

He glanced around, distracted by my outfit. I always wore it when the chron-senser networks said there was a jaunt about to happen. Their old dress styles were classic, so they weren't prepared for my peekaboo leggings, augmented breasts and perfectly symmetric face. The lipstick was outrageous for our time, but fit right into the twenty-first century kink.

He raised a flat ceramic thing and it whirred. Taking pictures, like the rest. They still hadn't learned, whenever this guy came from.

"Your pictures won't develop," I told him with a seemingly sympathetic smile.

"Huh? They gave me this —"

"You've heard of time paradoxes, yes? Space-time resolves those nicely. You can't take back knowledge that alters the past. All that gets erased automatically, a kind of information cleansing. Very convenient physics."

Startled, he glanced at his compact camera. "So … it'll be blank?"

"Yes," I said crisply. My left eye told me the chron-senser network was picking up an approaching closure. I leaned over and kissed him on the mouth. "Thanks! It's such a thrill to meet someone from the ancient times."

That shook him even more. Best to keep them off balance.

"So how do I get that cancer thing?" he said, eyes squinting with a canny cast.

"Let me have your clothes," I shot back.

"What? You want me … naked?"

"I can use them as antiques. That cancer stick is pretty expensive, so I'm giving you a good deal."

He nodded and started shucking off his coat, pants, shoes, wallet, coins, cash, set of keys. Reached for his shorts —

"Never mind the underwear."

"Oh." He handed me the bundle and I gave him the cancer stick. "Hey, thanks. I'll be back. We just wanted to see if —"

Pop. He vanished. The cancer stick rattled on the ground. It was just a prop, of course. Cancer was even worse now.

They never caught on. Of course, they don't have much time. That made the fifth this month, from several different centuries.

Time was like a river, yes. Go with the flow, it's easy. Fight against the current and space-time strips you of everything you're carrying back — pictures, cancer stick, memories. He would show up not recalling a thing. Just like the thousands of others I have turned into a nifty little sideline.

The past never seemed to catch on. Still, they stimulated interest in those centuries where time jaunters kept hammering against the laws of physics, like demented moths around a light bulb.

I hefted the clothes and wallet. These were in decent condition, grade 0.8 at least. They should fetch a pretty price. Good; I needed to eat soon. Time paid off, after all. A sucker born every minute, and so many, many moments in the rich past. ■

Gregory Benford is a physicist and a novelist. His best known novel is *Timescape*.